УДК 004.09 : 004.62 : 510.6
DOI: 10.15587/1729-4061.2018.150805

Підвищення ефективності забезпечення групової анонімності даних шляхом розроблення інформаційної технології

О. Р. Чертов, Д. Ю. Тавров

Широке впровадження в галузі офіційної статистики методів, що забезпечують анонімність даних про окремі групи (колективи) респондентів, стримується відсутністю відповідних промислових інформаційних технологій та систем. Запропоновано трирівневу клієнт-серверну архітектуру інформаційної технології забезпечення групової анонімності даних, у якій виділено клієнтів, сервери застосунків та бази даних, об'єднані в локальну мережу для підвищення безпеки первинних даних. Описано концептуальну модель даних у вигляді реляційної бази даних, наведено її ключові фрагменти. Дана модель охоплює всі основні сутності процесу забезпечення групової анонімності. Розглянуто реалізацію технології на основі платформи Java Enterprise Edition 8, сервера застосунків Oracle GlassFish Server, сервера баз даних MySQL та системи інженерних розрахунків SciLab.

Інформаційна технологія дає змогу забезпечувати групову анонімність даних у випадку існування загрози її порушення за рахунок аналізу даних допоміжного мікрофайлу. У технології передбачені операції побудови нечітких моделей груп за допомогою генетичного алгоритму та модифікація мікрофайлу за допомогою міметичного алгоритму, що дає змогу ефективно забезпечувати анонімність, уносячи в дані незначні спотворення. Загалом, запропонована інформаційна технологія базується на використанні шести застосунків: починаючи зі створення цільового подання мікрофайлу та завершуючи розв'язанням, власне, задачі забезпечення групової анонімності даних у мікрофайлі.

Застосування технології проілюстровано розв'язанням задачі забезпечення анонімності групи військових на основі реальних даних Спостереження за американським суспільством 2013 р. (American Community Survey – 2013). Показано, що розв'язання задачі силами колективу з п'яти фахівців дає змогу, щонайменше, в два з половиною рази пришвидшити процес підготовки мікрофайлу порівняно з існуючою технологією

Ключові слова: інформаційна технологія, групова анонімність, мікрофайл, нечітка модель, еволюційний алгоритм

1. Вступ

У сучасному світі спостерігається невпинне зростання обсягів цифрових даних, значна частка яких потребує оприлюднення з метою уможливлення проведення досліджень різного роду. При цьому повинен забезпечуватися належний захист даних від порушення приватності. Організації, які публікують дані з одночасним забезпеченням приватності, називають організаціями – розпоряд-

никами даних [1]. До організацій – розпорядників даних належать національні статистичні організації (наприклад, Державна служба статистики України), міжнаціональні статистичні організації (наприклад, Статистичний офіс Європейського Союзу), торгівельні асоціації, медичні заклади, бібліотеки, архіви тощо.

У своїх діяльності організації – розпорядники даних реалізують CSID-процес обробки даних, який складається з чотирьох підпроцесів – збір, зберігання, інтеграція, поширення (Capture, Storage, Integration, Dissemination). Дані збирають шляхом спостережень, переписів чи опитувань і зберігають у вигляді баз даних чи окремих файлів мікроданих (мікрофайлів). Поширення таких даних передбачає створення вихідних таблиць або мікрофайлів певної вибірки даних. Задачі, які організації – розпорядники даних розв'язують в рамках CSID-процесу обробки даних, включають уведення даних у базу та їх контроль, верифікацію даних, знеособлення даних, агрегацію даних, контроль над розкриттям даних, тобто їх анонімізацію.

Анонімність деякого об'єкта в множині даних – це властивість об'єкта бути нерозрізненним з-поміж інших елементів цієї множини [2]. Зазвичай, виділяють два види анонімності – індивідуальну, тобто пов'язану з інформацією про окремого респондента, та групову, тобто пов'язану з інформацією про групу осіб. Уперше забезпечення групової анонімності як складову CSID-процесу обробки даних розглянуто в [3].

Широке впровадження методів, що забезпечують анонімність даних про окремі групи респондентів, стримується відсутністю відповідних промислових інформаційних технологій та систем. Тому актуальною є задача розроблення інформаційної технології (ІТ) забезпечення групової анонімності даних як складової реалізації CSID-процесу обробки даних, яка повинна підвищувати ефективність процесу підготовки мікрофайлів до опублікування. Така ІТ повинна давати змогу будувати групу респондентів у мікрофайлі та забезпечувати її анонімність, найменшим чином викривлюючи при цьому дані мікрофайлу.

Оскільки виконання групової анонімізації потребує від користувачів ІТ різної підготовки, доцільно розподілити різні операції та дії між користувачами з різними ролями. Пропонується такий розподіл ролей:

– статистик, в обов'язки якого входить збір даних, їх уведення в базу даних (БД) тощо;

– датолог, в обов'язки якого входить підготовка метаданих про мікрофайли (опис їхньої структури) та їх модифікація;

– молодший аналітик, в обов'язки якого входить безпосередня анонімізація даних, у тому числі вибір параметрів групи та методів забезпечення групової анонімності, оцінювання якості одержуваних розв'язків;

– старший аналітик, в обов'язки якого входить ухвалення остаточного рішення про факт забезпечення групової анонімності;

– адміністратор БД, в обов'язки якого входить підтримка БД та підготовка мікрофайлів до опублікування.

Розроблювана ІТ повинна забезпечувати високий рівень надійності та безпеки первинних даних, а засоби розробки, використовувані на етапі її створення, повинні поширюватися у вільному доступі.

2. Аналіз літературних даних та постановка проблеми

Мікрофайл – це двовимірний масив даних, кожний рядок якого відповідає певному *респонденту* (*запису*), кожний стовпець – деякому *атрибуту* цього респондента (запису). Серед атрибутів виділяють три класи атрибутів:

– *параметризуючий* – атрибут, значення якого дають змогу розбити мікрофайл на *параметричні підмікрофайли*, тобто масиви даних, у яких записи мають однакове значення параметризуючого атрибута;

– *сутнісні* – атрибути, на основі значень яких можна сформувати критерії належності записів мікрофайлу певній групі. Називатимемо *моделлю* групи таку множину, яка складається з записів мікрофайлу з певними значеннями сутнісних та параметризуючого атрибутів мікрофайлу, що відповідають цій групі;

– *базові* – атрибути, які не є параметризуючими або сутнісними.

Для порушення групової анонімності на основі даних мікрофайлу будують *цільове подання мікрофайлу* (ЦПМ) відносно заданої групи. Найбільш поширеним на практиці є ЦПМ у вигляді *кількісного сигналу* – вектору значень, кожне з яких відповідає кількості респондентів, що належать групі, та мають значення параметризуючого атрибута, яке відповідає групі. Загроза порушення групової анонімності визначається в літературі [4] як можливість виявлення в ЦПМ *викидів* – значень, які суттєво перевищують за величиною решту значень. Викиди в ЦПМ свідчать про аномальну кількість респондентів, які належать групі, відносно деякого значення параметризуючого атрибута (наприклад, аномальна кількість військових у деякому регіоні).

Забезпечити групову анонімність можна тільки шляхом модифікації первинних даних, яка вносить у дані (бажано незначні) спотворення. Вилучення сутнісного атрибута з мікрофайлу на перший погляд є такою модифікацією, однак, як було показано в [5], у низці випадків існує можливість порушити групову анонімність, використовуючи сторонні дані. У [6] запропоновано метод побудови *нечіткої моделі групи респондентів* на основі *допоміжного мікрофайлу* – мікрофайлу, близького за структурою до того, для якого потрібно забезпечити анонімність. Така модель є множиною нечітких правил, за допомогою яких можна збудувати *допоміжне цільове подання* відносно допоміжного мікрофайлу (ДЦПМ), викиди в якому можуть збігатися з викидами ЦПМ. У випадку успішності побудови такої нечіткої моделі вилучення сутнісного атрибута з мікрофайлу не є гарантією забезпечення групової анонімності, і потрібно застосовувати додаткові методи її забезпечення.

Оскільки автоматизована побудова бази нечітких правил є складною задачею, у літературі для її розв’язання запропоновано застосовувати генетичні алгоритми [7]. Уперше такі алгоритми були запропоновані для побудови систем класифікаторів, що навчаються [8]. У літературі виділяють два основні підходи до побудови баз правил:

– згідно з мічиганським підходом [9], кожна особина в генетичному алгоритмі є окремим правилом;

– згідно з піттсбурзьким підходом [10], кожна особина в генетичному алгоритмі є повною базою правил.

Перевага мічиганського підходу полягає в тому, що [11] правила не залежать одне від одного, а його обчислювальна складність суттєво менша [12].

На сьогоднішній день найпотужнішим програмним продуктом для забезпечення анонімності даних є розроблений мовою Java безкоштовний пакет прикладних програм μ-ARGUS [13]. За допомогою μ-ARGUS можна забезпечувати індивідуальну анонімність мікрофайлів, використовуючи такі алгоритми, як перекодування та огрублення даних [14], *k*-анонімності [15], обміну даними [16], зашумлення [17].

У той же час, пакет μ-ARGUS не дає змоги забезпечувати групову анонімність, а дані мікрофайлів повинні зберігатися як окремі файли і не бути записані до бази даних, що спростило б побудову відповідної інформаційної технології.

Написаний мовою програмування *R* безкоштовний пакет прикладних програм sdcMicro [18] реалізує методи мікроагрегації [19], зашумлення, обміну даними, генерації синтетичних даних, перекодування та огрублення. Разом із тим, потреба у володінні середовищем *R* та в поданні мікрофайлів як окремих файлів звужує коло застосування цього програмного продукту. Пакет також не підтримує методи забезпечення групової анонімності.

У літературі описана інформаційна технологія [20–22], яка підтримує забезпечення групової анонімності даних. Мікрофайли за цієї технології зберігаються не як окремі файли, а в базі даних, однак при цьому

– не враховуються значення базових атрибутів, тобто відсутній захист від порушення групової анонімності шляхом аналізу допоміжних мікрофайлів;

– від користувача вимагається неодноразове застосування методу анонімізації з різними параметрами, поки не буде одержано мікрофайл задовільної якості;

– потрібен додатковий інструктаж для користувачів технології, оскільки в ній не передбачено розподіл ролей між користувачами з різними спеціалізаціями.

Тому є підстави вважати, що наразі фактично відсутні промислові інформаційні технології забезпечення групової анонімності даних, які б враховували комбінації значень базових атрибутів мікрофайлу і задовольняли б раніше висунуті вимоги. Це обумовлює потребу в розробленні такої інформаційної технології.

3. Мета та задачі дослідження

Метою роботи є підвищення ефективності забезпечення групової анонімності даних на етапі підготовки мікрофайлів до опублікування шляхом розроблення спеціалізованої інформаційної технології, яка дає змогу аналізувати дані допоміжних мікрофайлів.

Для досягнення цієї мети було розв’язано такі задачі:

– розробити архітектуру ІТ, яка задовольняє висунуті вище вимоги;

– розробити концептуальну модель БД для ІТ, яка містить усі потрібні сутності та відображає зв’язки між ними;

– реалізувати ІТ з використанням сучасним засобів розробки програмного забезпечення, які задовольняють висунуті вище вимоги;

– оцінити на практиці підвищення ефективності процесу підготовки мікрофайлів до опублікування з допомогою розробленої ІТ.

4. Матеріали та методи досліджень впливу розроблення інформаційної технології на ефективність підготовки мікрофайлів

4. 1. Задача забезпечення групової анонімності та методи її розв'язання

Позначимо через **M** мікрофайл, для якого потрібно забезпечити групову анонімність даних. Записи мікрофайлу позначимо через $\mathbf{r}^{(i)}$, i=1,…, ρ, атрибути – через $\mathbf{w}_j$, j=1,…, η. Кількість значень параметризуючого атрибута $\mathbf{w}_p$ позначимо через l_p. Тоді мікрофайл **M** можна розділити на параметричні підмікрофайли $\mathbf{M}_1$,…, $\mathbf{M}_{lp}$, у кожному з яких міститься ρ_i кількість записів. ЦПМ позначимо через $\mathbf{q}=(q_1, q_2,\ldots, q_{lp})$, де q_k – кількість сутнісних записів, які містяться в $\mathbf{M}_k$.

Індекси значень ЦПМ, які є викидами, позначимо через $OUT(\mathbf{q})$. Викиди в літературі [4] визначають за допомогою модифікованого методу τ Томпсона (ММТТ):

1. Знайти медіану $M_{\mathbf{q}}$ та псевдосередньоквадратичне відхилення $s_{ps\mathbf{q}}$ ЦПМ, значення якого впорядковано за зростанням:

$$M_q = \begin{cases} q_{(l_p+1)/2}, & l_p - \text{непарне}, \\ \dfrac{q_{l_p/2} + q_{l_p/2+1}}{2}, & l_p - \text{парне}, \end{cases}$$

$$s_{ps\mathbf{q}} = \frac{q_{0,75} - q_{0,25}}{1,349},$$

де $q_{0,75}$ та $q_{0,25}$ – верхня та нижня квартилі, відповідно.

2. Обчислити абсолютні відхилення від медіани $\forall q_i$, i=1,…, l_p,:

$$d_i = \left| q_i - M_{\mathbf{q}} \right|.$$

3. Обчислити величину

$$\tau = \frac{t_{\alpha/2} \cdot \left(l_p - 1\right)}{\sqrt{l_p}\sqrt{l_p - 2 + t^2_{\alpha/2}}},$$

де $t_{\alpha/2}$ – критичне значення t розподілу Ст'юдента для кількості ступенів свободи (l_p – 2) та рівня значущості α.

4. Якщо $d_i > \tau s_{ps\mathbf{q}}$, то i-е значення ЦПМ є викидом. Тоді його вилучають із ЦПМ та переходять на крок 1. Якщо для жодного i цей критерій не виконується, алгоритм завершується.

Задача забезпечення групової анонімності (ЗЗГА) полягає в підборі такої модифікації даних, що в ЦПМ $\mathbf{q}^*$, побудованому для модифікованого мікрофайлу $\mathbf{M}^*$, викиди, обчислені за ММТТ, масковано, тобто $OUT(\mathbf{q}) \cap OUT(\mathbf{q}^*) = \emptyset$.

При цьому спотворення, внесені в дані мікрофайлу, повинні бути незначні. На практиці мікрофайли, зазвичай, модифікують шляхом попарного обміну респондентів, схожих у розумінні визначальної метрики [22]:

$$\mathrm{InfM}(\mathbf{r}^{(i)},\mathbf{r}^{(j)}) = \sum_{k=1}^{n_{\mathrm{cat}}} \gamma_k \chi^2\left(r_{I_k}^{(i)}, r_{I_k}^{(j)}\right) + \sum_{l=1}^{n_{\mathrm{ord}}} \omega_l \left(\frac{r_{J_l}^{(i)} - r_{J_l}^{(j)}}{r_{J_l}^{(i)} + r_{J_l}^{(j)}}\right)^2, \qquad (1)$$

де I_k (J_l) – k-ий категорійний (l-ий порядковий) визначальний атрибут, тобто атрибут, цікавий для потенційних дослідників мікрофайлу; $\chi(v_1, v_2)$ дорівнює числу χ_1, якщо значення v_1 та v_2 атрибутів належать одній категорії, та χ_2 – інакше; γ_k та ω_l – невід'ємні вагові коефіцієнти (що більша вага, то важливіший атрибут).

Оскільки вибір конкретного $\mathbf{q}^*$ асоціюється з конкретним обсягом спотворень, обчисленим як сума значень (1) для кожної пари респондентів, ЗЗГА зводиться до підбору $\mathbf{q}^*$, який забезпечить мінімальний обсяг спотворень [22]. Заздалегідь визначити оптимальний $\mathbf{q}^*$ неможливо, тому при розв'язанні ЗЗГА на значення $\mathbf{q}$ накладають *нечіткі обмеження* [23], що задаються функціями $\mu_i(x)$ для кожного i-ого значення $\mathbf{q}$. Кожна така функція дорівнює 1 для $x \le \varepsilon_j$, монотонно спадає до 0 для $\varepsilon_j < x \le q_j$, та дорівнює 0 для $x > q_j$, де ε_j – порогове значення, нижче якого повинно зменшитися i-те значення ЦПМ $\mathbf{q}$. Величину $\mu_i(q_i)$ називають *сумісністю* q_i з накладеним на нього нечітким обмеженням. Сумісність $\mu(\mathbf{q})$ усього сигналу $\mathbf{q}$ з набором нечітких обмежень визначають як добуток усіх $\mu_i(q_i)$, $i=1,\ldots, l_p$. Порогові значення доцільно встановлювати як

$$\varepsilon_j = \mathbf{q}^{K\max} - (q_j - \mathbf{q}^{K\max}) \cdot 0,2, \qquad (2)$$

де $\mathbf{q}^{K\max}$ – K-е найбільше значення підмножини значень ЦПМ $\mathbf{q}'=(q_j)$, яка складається зі значень з індексами, що належать $OUT'(\mathbf{q})$, – доповненню $OUT(\mathbf{q})$ до множини $\{1,\ldots, l_p\}$.

З урахуванням вищевказаного, ЗЗГА можна сформулювати як задачу пошуку послідовності попарних обмінів записів у вигляді

$$\mathbf{S} = \left(\left(\mathbf{r}^{(i_1)}, \mathbf{r}^{(j_1)}\right), \ldots, \left(\mathbf{r}^{(i_Q)}, \mathbf{r}^{(j_Q)}\right)\right), \qquad (3)$$

де i_k, j_k, $k=1,\ldots, Q$ – індекси записів мікрофайлу, які обмінюються між підмікрофайлами в рамках модифікації. Кожна така послідовність, яку називатимемо *розв'язком* ЗЗГА, повинна задовольняти такі умови:

$$\mu\left(\mathbf{q}^*(\mathbf{S})\right) \ge \alpha_{comp},$$

$$\frac{\left|OUT(\mathbf{q}) \cap OUT\left(\mathbf{q}^*(\mathbf{S})\right)\right|}{\left|OUT(\mathbf{q})\right|} \le K_{out},$$

$$\sum_{k=1}^{Q} \mathrm{InfM}\left(\mathbf{r}^{(i_k)}, \mathbf{r}^{(j_k)}\right) \le K_{dist} \cdot C_{\max}, \qquad (4)$$

де $\mathbf{q}^*(\mathbf{S})$ – модифіковане ЦПМ; $\mu(\mathbf{q}^*(\mathbf{S}))$ – сумісність $\mathbf{q}^*$ з накладеними нечіткими обмеженнями; α_{comp} – поріг сумісності; K_{out} – поріг чутливості; K_{dist} – поріг спотворень; $C_{\max}$ – максимальне сумарне значення (1) для розв'язуваної ЗЗГА.

У літературі пошук послідовностей зазначеного вище формату здійснюють за допомогою *міметичних алгоритмів* (МА) – еволюційних алгоритмів, які поєднують стохастичність із елементами локального пошуку [24]. Популяція в МА для розв'язання ЗЗГА складається з матриць розмірності $Q \times 4$, які позначаються через U. Кожний рядок матриці визначає записи для обміну між підмікрофайлами у такий спосіб:

– елемент u_{i1}, $i=1,\ldots, Q$ – індекс підмікрофайлу, із якого потрібно вилучити запис; елемент u_{i3}, $i=1,\ldots, Q$ – індекс підмікрофайлу, у який потрібно додати запис;

– елемент u_{i2}, $i=1,\ldots, Q$ – індекс запису в рамках підмікрофайлу, який потрібно вилучити; елемент u_{i4}, $i=1,\ldots, Q$ – індекс запису в рамках підмікрофайлу, який потрібно обміняти на запис, визначений u_{i1} та u_{i2}.

На структуру U накладаються певні обмеження. Зокрема, кількість входжень індексу підмікрофайлу i, $i=1,\ldots, l_p$, у перший стовпець не може перевищувати q_i, а у третій – $(\rho_i - q_i)$. Записи мікрофайлу не можуть зустрічатися в U більше одного разу.

Функція пристосованості особин у популяції має вигляд:

$$f(U) = \Phi(U) \cdot \Upsilon(U) \cdot \Psi(U) =$$

$$= \prod_{j=i_1}^{i_k} \mu_{A_j}\left(q_j^*(U)\right) \times \frac{C_{\max} - \sum_{i=1}^{Q} \mathrm{InfM}\left(\mathbf{M}_{u_{i1}}(u_{i2}), \mathbf{M}_{u_{i3}}(u_{i4})\right)}{C_{\max}} \times \frac{1}{1 + e^{\frac{1}{2}(Q_U - L)}}, \qquad (5)$$

де $\Phi(U)$ – сумісність $\mu(\mathbf{q})$ сигналу з накладеними нечіткими обмеженнями (міра якості маскування викидів на проміжку [0, 1]); $\Upsilon(U)$ – міра якості мінімізації обсягу спотворень на проміжку [0, 1]; $\Psi(U)$ – штрафний терм (на проміжку [0, 1]) проти необмеженого збільшення розмірності особин.

Міметичний алгоритм складається з таких кроків:

1. Згенерувати популяцію $P=\{U_i\}$ з μ особин, $i=1,\ldots, \mu$, застосувати до кожної з них оператор локального пошуку $LS(U_i)$, $i=1,\ldots, \mu$.

2. Обчислити значення функції пристосованості (5) $\forall U_i$ з P.

3. Перевірити умову завершення та зупинити алгоритм, якщо вона виконується.

4. Відібрати λ пар батьківських особин та помістити їх у множину P'.

5. Застосувати оператор схрещування $REC(U_{i1}, U_{i2})$ до кожної пари U_{i1}, U_{i2} з P'. Помістити нащадків у множину P''.

6. Застосувати оператор мутації $MUT(U_j)=(MUT_4 \circ MUT_3 \circ MUT_2 \circ MUT_1)(U_j)$ $\forall U_j$ з P''. Кожний оператор MUT_k, k=1, …, 4, діє на k-ий стовпець особини U_j окремо.

7. Застосувати $LS(U_j)$ $\forall U_j$ з P''.

8. Обчислити значення функції пристосованості (5) $\forall U_i$ з P''.

9. Відібрати μ особин з $P \cup P''$ із найбільшим значенням функції пристосованості та помістити у P на місце μ особин із найменшим значенням функції пристосованості.

10. Перейти на крок 2.

Як оператор схрещування в МА використовується оператор розрізання [25], як оператори мутації – мутації обміну (MUT_1, MUT_3) [26] та мутації випадкової заміни (MUT_2, MUT_4) [27], як оператор відбору – оператор турнірного відбору [28].

Оператор локального пошуку [22] передбачає виконання таких кроків:

1. Для кожного рядка з U виконати кроки 2–3.

2. Випадково згенерувати число r, рівномірно розподілене на [0, 1].

3. Якщо $r \le p_{mem}$ ($r > p_{mem}$), присвоїти елементу u_{i4} (u_{i2}) індекс запису з $\mathbf{M}_{ui3}$ ($\mathbf{M}_{ui1}$), найближчий у розумінні (1) до запису u_{i2} (u_{i4}) з $\mathbf{M}_{ui1}$ ($\mathbf{M}_{ui3}$).

Початкову популяцію генерують випадковим чином, при цьому особини повинні мати різну кількість рядків. Критерієм завершення, зазвичай, слугує кількість виконаних поколінь алгоритму.

Як було зазначено вище, інколи існує можливість порушити групову анонімність, навіть якщо знайдено розв'язки, які задовольняють вимоги, висунуті вище. Маючи доступ до допоміжного мікрофайлу $\mathbf{M}^{aux}$, можна збудувати *нечітку модель групи*, на основі якої можна визначити викиди ЦПМ в **M**. Відповідний процес складається з таких кроків [6]:

– за допомогою такої нечіткої моделі кожному запису мікрофайлу можна зіставити ступінь його належності групі $\mu_G(\mathbf{r}^{(i)})$ як значення з проміжку [0, 1]. Ступінь належності є мірою достовірності належності запису групі в умовах відсутності сутнісних атрибутів, які однозначно вказують на таку належність;

– на основі ступенів належності усіх записів можна збудувати ДЦПМ у вигляді

$$q_j^{aux} = \left| \mathbf{r} \in \mathbf{M}_j \mid \mu_G(\mathbf{r}) \ge \alpha \right|, \quad j = 1,...,l_p, \tag{6}$$

де α – поріг належності, використовуючи який, можна відсіяти записи з низьким ступенем належності (як правило, беруть α=0,5);

– виявити за допомогою ММТТ викиди в (6).

Для уможливлення побудови нечіткої моделі допоміжний мікрофайл $\mathbf{M}^{aux}$ повинен бути близьким за структурою до основного мікрофайлу **M**. Зокрема, два мікрофайли можна гармонізувати, тобто привести до єдиної структури, з атрибутами, значення яких мають однакову інтерпретацію.

Нечітка модель групи складається з умовних нечітких висловлювань (нечітких правил) R_i, i=1,… , m, канонічна форма яких має вигляд

$$R_i: \text{ Якщо } L_1 \in A_{i1}, L_2 \in A_{i2}, ..., L_t \in A_{it}, \text{ то } G, \tag{7}$$

де L_k, k=1,…, t – лінгвістичні змінні [29], базові змінні яких визначені на множинах значень базових атрибутів мікрофайлу w_{bk}, k=1,…, t, відповідно; A_{ij} – нечітке значення змінної L_j, яке зустрічається в нечіткому правилі R_i; G – клас записів, які належать групі. Логічна зв'язка «і» моделюється як нечіткий перетин у вигляді добутку.

Кожна лінгвістична змінна L_k у нечіткій моделі групи відповідає деякому базовому атрибуту мікрофайлу w_{bk}, k=1,…, t. При цьому кожна змінна L_k, k=1,…, t, має декілька значень LL_k^j, j=1,…, l_{Lk}. Записи, значення яких не належать носію хоча б однієї базової змінної відповідної лінгвістичної змінної, вилучаються з мікрофайлу.

Задача формування множини нечітких правил вигляду (7) може розглядатися як задача виявлення підгруп [30] у множині респондентів, розподіл яких становить цікавість для досліджень. Із кожним правилом з нечіткої моделі асоційована міра якості [31], яка вказує на його здатність ефективно виявляти цікаву підгрупу. У даній роботі використовуються міри якості, запропоновані в [6]:

– фактор дискримінації:

$$DF(R_i) = \frac{\sum\limits_{\mathbf{r}^{aux} \in G} APC^{\alpha}(\mathbf{r}^{aux}, R_i)}{\rho_v^{aux}} - \frac{\sum\limits_{\mathbf{r}^{aux} \in \mathbf{M}^{aux}} APC^{\alpha}(\mathbf{r}^{aux}, R_i)}{\rho^{aux}}, \tag{8}$$

де ρ^{aux} – кількість записів у $\mathbf{M}^{aux}$; $\rho^{aux}{}_v$ – кількість сутнісних записів у $\mathbf{M}^{aux}$; $\mathbf{r}^{aux}$ – запис із $\mathbf{M}^{aux}$; APC^{α} – сумісність запису з антецедентом нечіткого правила:

$$APC^{\alpha}(\mathbf{r}, R_i) = \begin{cases} \prod\limits_j \mu_{A_{ij}}(r_{b_j}), & \text{якщо } \prod\limits_j \mu_{A_{ij}}(r_{b_j}) \geq \alpha, \\ 0, & \text{інакше,} \end{cases}$$

де μ_{Aij} – функція належності значення A_{ij} лінгвістичної змінної L_j, яке міститься в правилі R_i; П – нечіткий перетин; r_{bj} – значення j-го базового атрибута запису $\mathbf{r}$. Додатне значення (8) свідчить про те, що нечітке правило відносить до групи непропорційне більше сутнісних записів із $\mathbf{M}^{aux}$, ніж записів із $\mathbf{M}^{aux}$ загалом;

– фактор відносної достовірності:

$$RCF(R_i) = \frac{\sum\limits_{\mathbf{r}^{aux} \in G} APC^{\alpha}(\mathbf{r}^{aux}, R_i)}{\sum\limits_{\mathbf{r}^{aux} \notin G} APC^{\alpha}(\mathbf{r}^{aux}, R_i)}. \tag{9}$$

Значення (9), що перевищує поріг відносної достовірності γ, свідчить про те, що нечітке правило неправильно класифікує як належних групі невелику кількість записів із $\mathbf{M}^{aux}$.

Зменшуване в (8) називають носієм нечіткого правила та позначають через к.

Правила для нечіткої моделі групи можна збудувати автоматизовано на основі *генетичного алгоритму* (ГА), уперше запропонованого в [32], який відповідає описаному вище мічиганському підходу. Популяція в ГА для побудови нечітких правил складається з особин, кожна з яких відповідає окремому нечіткому правилу. Кожне правило подається у вигляді вектору $R_i=(R_{i1}, R_{i2},\ldots, R_{it})$, значення якого є індексами нечітких значень лінгвістичних змінних L_k, $k=1,\ldots, t$.

Функція пристосованості особин у популяції має вигляд:

$$f\left(R_i\right)=\begin{cases} DF\left(R_i\right)\cdot RCF\left(R_i\right), & DF\left(R_i\right)>0, \\ 0, & DF\left(R_i\right)\le 0, \end{cases} \quad i=1,2,\ldots,\mu. \tag{10}$$

Генетичний алгоритм складається з таких кроків:

1. Згенерувати популяцію $\mathbf{R}=\{R_i\}$ з μ правил, $i=1,\ldots, \mu$.
2. Обчислити значення функції пристосованості (10) $\forall R_i$ з **R**.
3. Перевірити умову завершення та зупинити алгоритм, якщо вона виконується.
4. Відібрати λ пар батьківських особин та помістити їх у множину **R′**.
5. Застосувати оператор схрещування $REC(R_{i1}, R_{i2})$ до кожної пари R_{i1}, R_{i2} з **R′**. Помістити нащадків у множину **R″**.
6. Застосувати оператор мутації $MUT(R_j)$ $\forall R_j$ з **R″**.
7. Обчислити значення функції пристосованості (10) $\forall R_i$ з **R″**.
8. Замінити λ пар особин з **R** із найменшим значенням функції пристосованості на особин з **R″**.
9. Перейти на крок 2.

Як оператор схрещування в ГА використовується оператор рівномірної рекомбінації [33], як оператор мутації – мутація випадкової заміни [27], як оператор відбору – оператор турнірного відбору [28]. Початкову популяцію генерують випадковим чином. Критерієм завершення, як правило, слугує кількість виконаних поколінь алгоритму.

Правила, одержувані в результаті застосування ГА, повинні задовольняти такі умови:

– мають додатне значення (8);

– мають значення (9), що перевищує наперед заданий поріг γ;

– мають носій к, що перевищує наперед задане значення;

– не є частинними випадками більш загальних правил.

Адекватність збудованої нечіткої моделі групи для задачі виявлення викидів у ЦПМ на основі допоміжного мікрофайлу оцінюватимемо за допомогою метрики, запропонованої в [34]:

$$MB = \min_{\substack{0 \le t_1 \le l_p \\ 0 \le t_2 \le l_p}} \ln B(t_1, t_2), \tag{11}$$

де B – Баєсівський фактор, що обчислюється як

$$B(t_1,t_2) = \left[\frac{TP + FP + FN + TN + 1}{(TP + FP + t_1 + 1)(FN + TN + t_2 + 1)} \right] \times$$

$$\times \left[\frac{(t_1 + 1)(t_2 + 1)}{t_1 + t_2 + 1} \right] \cdot \frac{(TP + FP + FN + TN)!}{(TP + FN)!(FP + TN)!} \times$$

$$\times \sum_{i=0}^{t_1} \sum_{j=0}^{t_2} \frac{\left(\frac{t_1!}{i!(t_1 - i)!} \right)^2 \times \left(\frac{t_2!}{j!(t_2 - j)!} \right)^2}{\frac{(t_1 + t_2)!}{(i + j)!(t_1 + t_2 - i - j)!} \times \frac{(TP + FP + t_1)!}{(TP + i)!(FP + t_1 - TP - i)!} \times \frac{(FN + TN + t_2)!}{(FN + j)!(TN + t_2 - FN - j)!}},$$

де t_1, t_2 – невід'ємні цілі числа; TP, FP, TN, FN – елементи матриці невідповідностей

$$\mathbf{Z} = \begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} =$$

$$= \begin{pmatrix} \left| OUT(\mathbf{q}) \cap OUT(\mathbf{q}^{aux}) \right| & \left| OUT'(\mathbf{q}) \cap OUT(\mathbf{q}^{aux}) \right| \\ \left| OUT(\mathbf{q}) \cap OUT'(\mathbf{q}^{aux}) \right| & \left| OUT'(\mathbf{q}) \cap OUT'(\mathbf{q}^{aux}) \right| \end{pmatrix}.$$

Інтерпретація значень (11) здійснюється за табл. 1.

Таблиця 1

Інтерпретація значень метрики адекватності нечіткої моделі групи

Значення метрики	Адекватність
Менше 0	Дуже низька
Від 0 до 1	Низька
Від 1 до 3	Середня
Від 3 до 5	Сильна
Більше 5	Дуже сильна

Зазвичай, п'яти різних значень такої метрики достатньо для якісного опису адекватності моделі.

4. 2. Архітектура інформаційної технології забезпечення групової анонімності та концептуальна модель даних

У даній роботі пропонується трирівнева клієнт-серверна архітектура ІТ (рис. 1), у якій виділяються АРМ користувачів у різних ролях, сервер застосунків та сервер БД. Такий підхід до побудови ІТ дає змогу виконати поставлені в розділі 1 вимоги, зокрема:

– підтримка користувачів у різних ролях;

– забезпечення високого рівня надійності та безпеки первинних даних за рахунок їх збереження на окремому сервері БД з обмеженим доступом;

– забезпечення високої гнучкості та ефективності системи шляхом розподілу задач між серверами застосунків та БД.

Складові ІТ виконують такі функції:

– сервер застосунків керує підключеннями та транзакціями клієнтів, їх аутентифікацією, паралельним виконанням потоків, балансуванням навантаження в мережі тощо;

– сервер БД керує базою даних, забезпечуючи цілісність даних, опрацьовує запити від клієнтів, керує обліковими записами користувачів тощо. Клієнти не мають прямого доступу до БД (усі комунікації здійснюються через сервер застосунків), що підвищує рівень безпеки даних;

– АРМ клієнтів надають можливість виконувати функції відповідно до розподілу обов’язків. Так, статистик може переглядати та редагувати параметри групи, викиди ЦПМ, читати метадані. Датолог може переглядати та редагувати метадані. Молодший аналітик може переглядати та редагувати значення ЦПМ, параметри нечітких моделей, ММТТ, ГА, МА, розв’язків ЗЗГА, переглядати метадані, параметри групи та ЗЗГА. Старший аналітик може переглядати будь-яку інформацію, редагувати параметри ЗЗГА. Адміністратор БД може переглядати будь-яку інформацію з БД, редагувати метадані.

Застосунки реалізують такі функції:

– застосунок створення ЦПМ будує відповідне ЦПМ чи допоміжне ЦПМ;

– застосунок гармонізації мікрофайлів гармонізує основний та допоміжний мікрофайли;

– застосунок виявлення викидів виявляє за допомогою ММТТ викиди в ЦПМ;

– застосунок побудови нечітких правил запускає ГА для побудови правил нечіткої моделі групи;

– застосунок верифікації адекватності моделі обчислює метрику адекватності (11) для заданої моделі;

– застосунок розв’язання ЗЗГА запускає МА із заданими параметрами для розв’язання відповідної задачі.

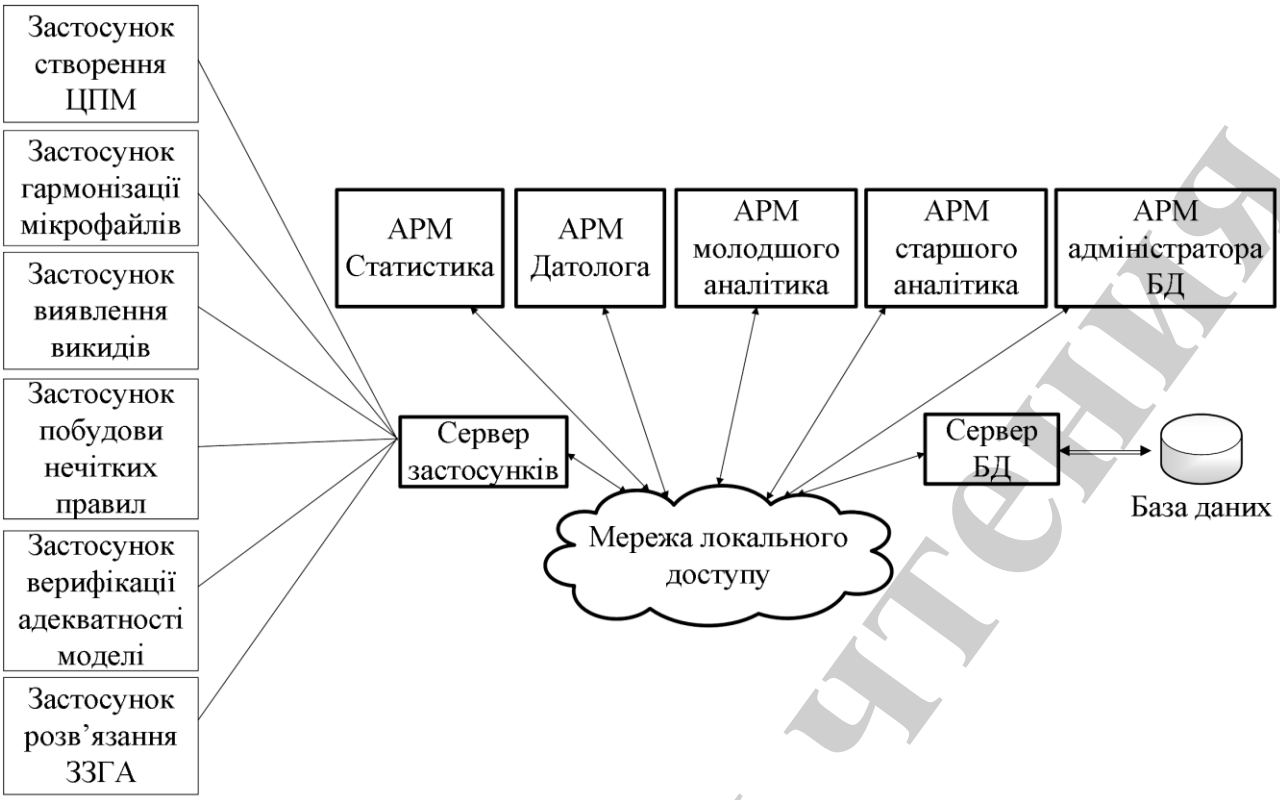

Рис. 1. Архітектура інформаційної технології забезпечення групової анонімності даних

Інформаційну технологію, що пропонується в даній роботі, можна подати у вигляді трьох етапів: етап побудови моделі групи (Е1), етап побудови нечіткої моделі групи (Е2), етап розв'язання ЗЗГА (Е3). Етап Е1 складається з такий операцій та дій:

– операція О1-1 — «Задання групи» (дії Д1-1-1 «Завантаження метаданих», Д1-1-2 «Задання атрибутів», Д1-1-3 «Задання типу задачі»);

– операція О1-2 — «Виявлення викидів у ЦПА» (дії Д1-2-1 «Побудова ЦПМ», Д1-2-2 «Виконання ММТТ», Д1-2-3 «Відсіювання викидів»).

Етап Е2 складається з такий операцій та дій:

– операція О2-1 — «Створення допоміжного мікрофайлу» (дії Д2-1-1 «Завантаження допоміжних метаданих», Д2-1-2 «Гармонізація мікрофайлів»);

– операція О2-2 — «Побудова нечіткої моделі» (дії Д2-2-1 «Задання базових атрибутів», Д2-2-2 «Визначення нечітких значень», Д2-2-3 «Задання параметрів та виконання ЕА»);

– операція О2-3 — «Верифікація адекватності нечіткої моделі» (дії Д2-3-1 «Побудова ДЦПМ», Д2-3-2 «Виконання ММТТ», Д2-3-3 «Відсіювання викидів», Д2-3-4 «Обчислення метрик адекватності»).

Етап Е3 складається з такий операцій та дій:

– операція О3-1 — «Задання параметрів ЗЗГА» (дії Д3-1-1 «Задання порогів», Д3-1-2 «Задання *K*-го найбільшого значення ЦПМ»);

– операція О3-2 — «Виконання МА» (дії Д3-2-1 «Задання функцій належності нечітких обмежень», Д3-2-2 «Задання параметрів МА», Д3-2-3 «Запуск МА», Д3-2-4 «Відбір розв'язку ЗЗГА»).

Послідовність виконання всіх дій клієнтами різних ролей наведено на UML-діаграмах діяльності (рис. 2, 3).

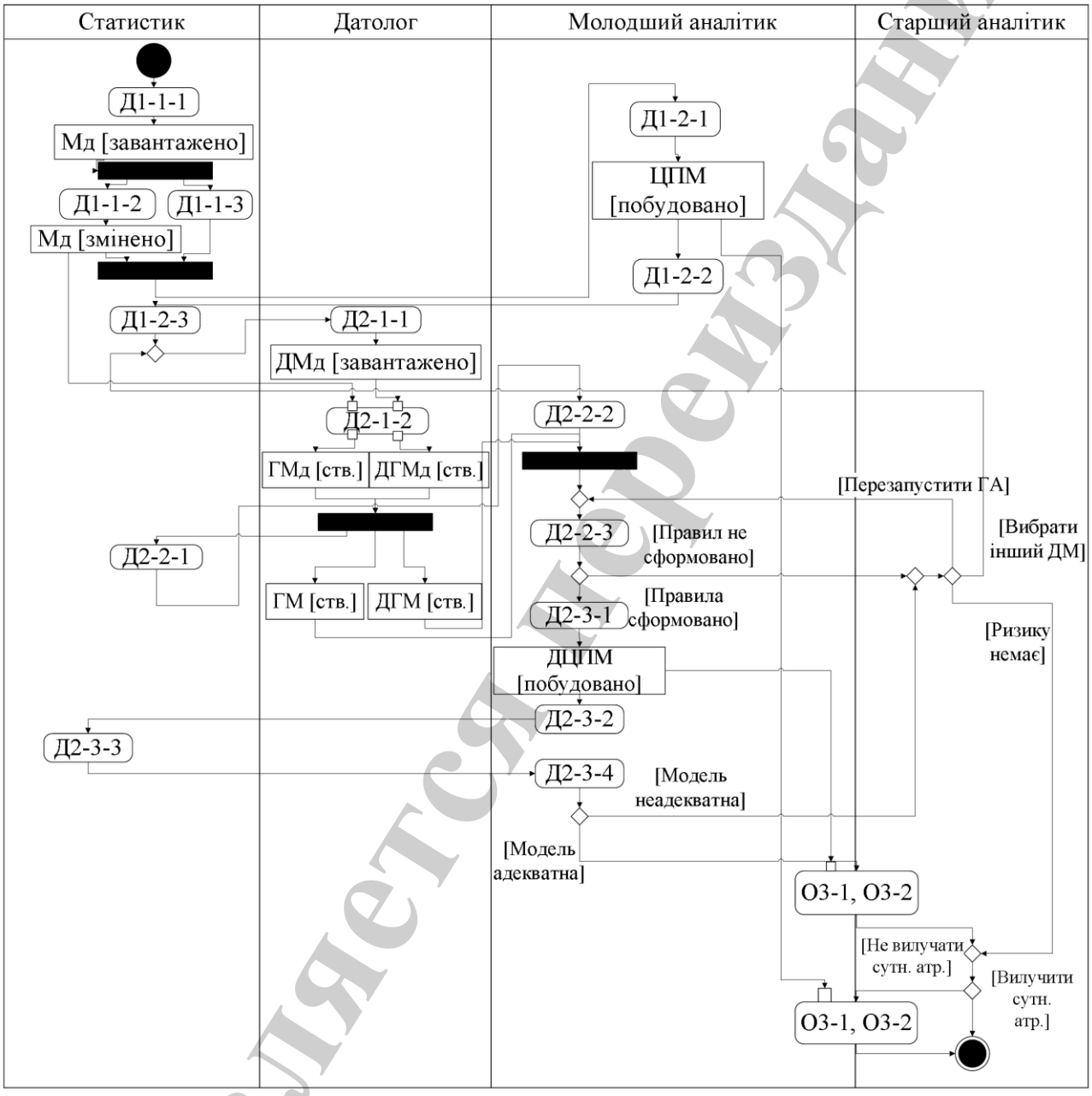

Рис. 2. Діаграма діяльності клієнтів ІТ забезпечення групової анонімності (етапи Е1, Е2)

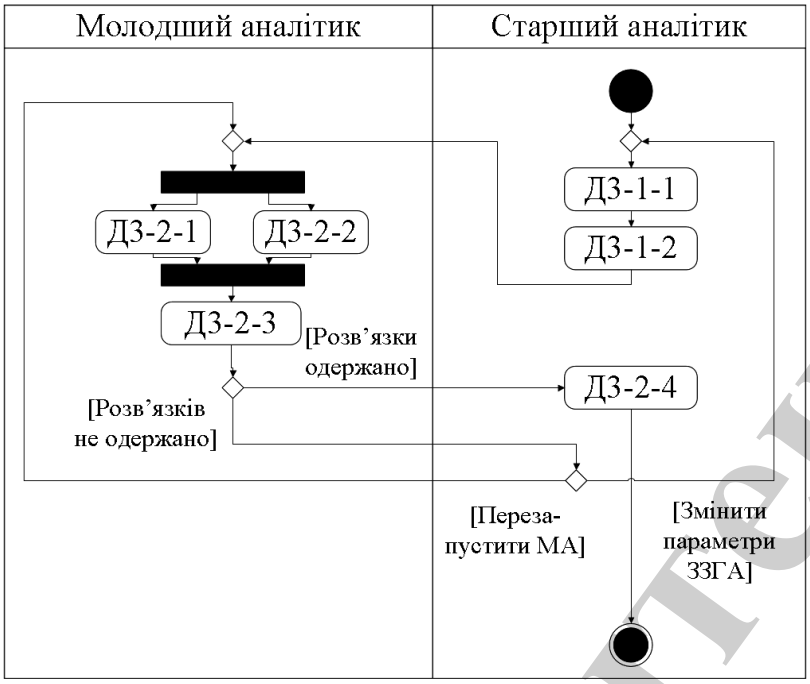

Рис. 3. Діаграма діяльності клієнтів ІТ забезпечення групової анонімності (етап Е3)

Концептуальну модель даних, розроблену в рамках запропонованої ІТ у вигляді реляційної БД, розташованої на сервері БД, можна подати у вигляді декількох фрагментів. На рис. 4 наведено фрагмент концептуальної моделі даних, який відповідає сутностям мікрофайлу та його цільового подання. Сутності, винесені в цей фрагмент, відповідають таким поняттям, пов'язаним із забезпеченням групової анонімності:

– сутність Microfile відповідає даним мікрофайлів та містить, крім первинного ключа ID_Microfile (ідентифікатор), обов'язкові атрибути MI_Name (назва мікрофайлу), MI_Desc (опис мікрофайлу), MI_Data (дані мікрофайлу у форматі BLOB);

– сутність Attribute відповідає атрибутам мікрофайлів та містить, крім первинного ключа AT_Name (назва атрибута), обов'язкові атрибути AT_Desc (опис атрибута) та AT_Type (тип атрибута: номінальний або категорійний);

– сутність AttrValue відповідає значенням атрибутів мікрофайлів та містить обов'язкові атрибути AV_Date (дата занесення значення), AV_Value (значення) та AV_Desc (семантичний опис значення);

– сутність AttrCharacteristic відповідає характеристикам атрибутів мікрофайлів та містить, окрім первинного ключа AC_Date (дата створення), обов'язкові атрибути AC_Weight (вага атрибута, яка використовується в метриці (1)) та AC_Xi (параметр χ з метрики (1));

– сутність GRM відповідає ЦПМ та містить, крім первинного ключа GR_Date (дата створення), обов'язковий атрибут GR_Data (значення ЦПМ у вигляді текстового рядку з розділеними комами значеннями);

– сутність MMTT_Params відповідає параметрам MMTT та містить, крім первинного ключа MMTT_Date (дата створення), обов'язковий атрибут MMTT_Alpha (рівень значущості α з MMTT, описаного у розділі 4.1);

– сутність MMTT_GRM відповідає викидам сигналу: необов'язковий атрибут GR_Outliers містить індекси значень ЦПМ, які відповідають викидам, виявленим з використанням MMTT у форматі, аналогічному формату зберігання значень самого ЦПМ;

– сутність Visual_Outlier відповідає викидам ЦПМ, відібраним статистиком для маскування, та містить, крім первинного ключа VO_Date (дата створення), обов'язковий атрибут VO_Outlier, який містить індекси значень ЦПМ, які відповідають цим викидам у форматі, аналогічному формату зберігання значень самого ЦПМ;

– сутність Problem відповідає ЗЗГА та містить, крім первинного ключа ID_Problem (ідентифікатор), обов'язкові атрибути PR_Date (дата створення задачі), PR_Stage (номер етапу, на якому перебуває задача) та необов'язковий атрибут PR_RemoveVital (прапорець, який вказує, чи потрібно вилучати сутнісні атрибути з мікрофайлу);

– сутність User відповідає користувачам ІТ та містить, крім первинного ключа ID_User (ідентифікатор), обов'язкові атрибути US_Login (логін користувача), US_Password (пароль користувача), US_Role (роль користувача), US_Name (ім'я користувача), US_IsActive (прапорець, який вказує на те, чи є користувач активним у системі) та US_Date (дата створення користувача);

– сутність Role відповідає ролям користувачів ІТ та містить, крім первинного ключа RO_Date (дата створення), обов'язкові атрибути RO_Title (назва ролі) та RO_IsActive (прапорець, який вказує на те, чи є роль активною в системі);

– сутність UserRole потрібна для організації зв'язку «багато-до-багатьох» між користувачами ІТ та їхніми ролями.

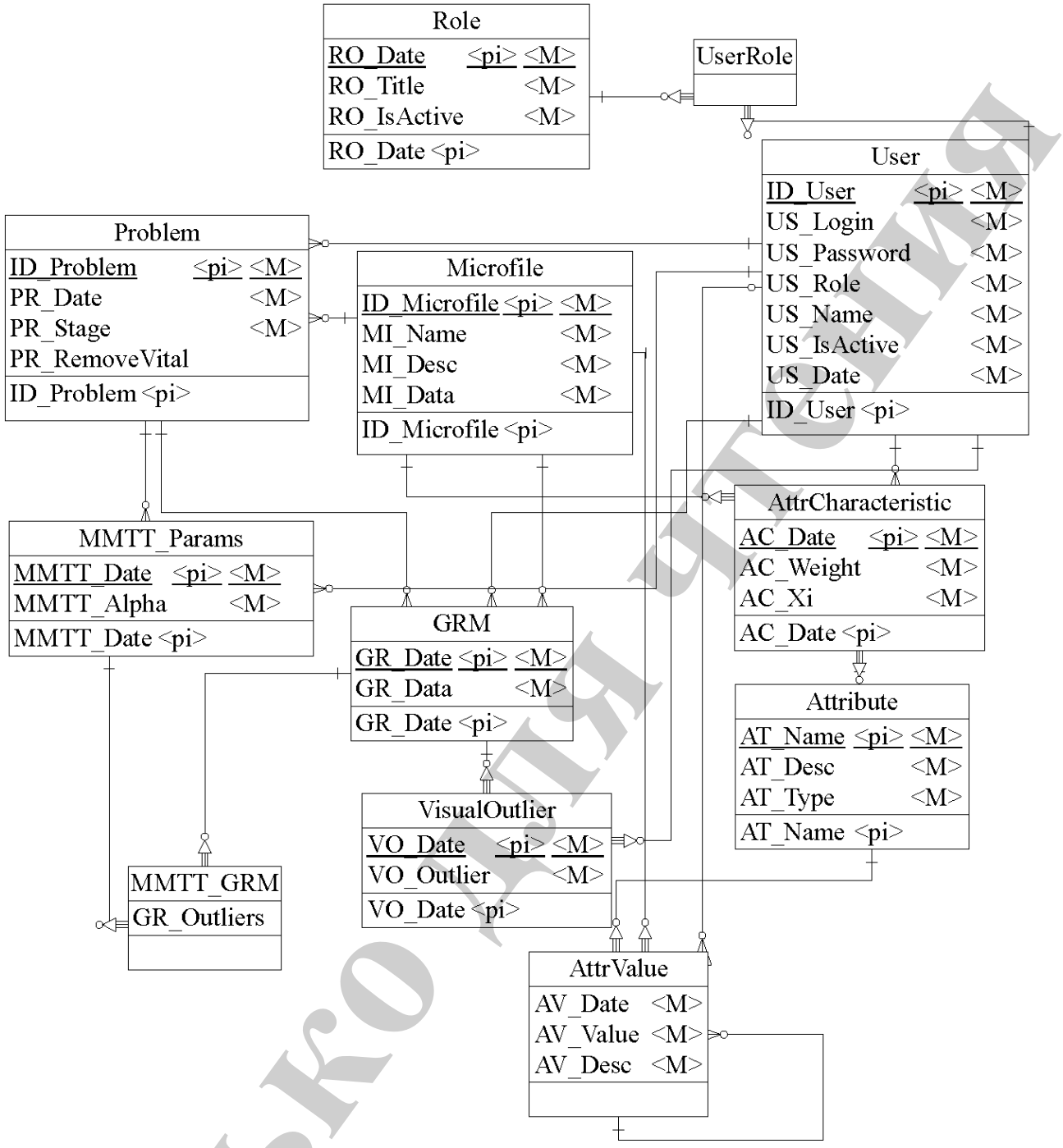

Рис. 4. Фрагмент концептуальної моделі даних, що відповідає мікрофайлу

На рис. 5 наведено фрагмент концептуальної моделі даних, який відповідає нечітким моделям груп у мікрофайлі. Сутності, винесені у цей фрагмент, відповідають таким поняттям, пов'язаним із забезпеченням групової анонімності:

– сутність FuzzyRule відповідає правилам нечіткої моделі групи та містить, окрім первинного ключа FRU_Date (дата створення правила), такі обов'язкові атрибути:

– FRU_Rule (правило у вигляді рядку з розділеними комами значеннями, і кожне з них відповідає індексу нечіткого значення лінгвістичної змінної з числа асоційованих із даним правилом);

– FRU_DF (фактор дискримінації (8) правила);

– FRU_RCF (фактор відносної достовірності (9) правила);

– FRU_Kappa (носій к правила);

– сутність FuzzyModelParameter відповідає нечітким значенням лінгвістичних змінних, що входять до складу моделі, та містить, крім первинного ключа FMP_Date (дата створення параметрів), обов'язкові атрибути FMP_Name (назва значення) та FMP_Params (параметри функції належності відповідного нечіткого значення у вигляді рядку з розділеними комами значеннями);

– сутність LinguisticVariable відповідає лінгвістичним змінним та містить, крім первинного ключа LV_Date (дата створення змінної), обов'язкові атрибути LV_Desc (опис змінної) та LV_Name (назва змінної);

– сутність ModelAdequacy відповідає значенням метрик адекватності нечіткої моделі групи та містить, крім первинного ключа MAD_Date (дата створення значень метрик), обов'язковий атрибут MAD_MB (значення метрики (11));

– сутність GA_Params відповідає параметрам ЕА для створення нечіткої моделі групи та містить, крім первинного ключа GA_Date (дата створення параметрів), такі обов'язкові атрибути:

– GA_Mu (розмір популяції) та GA_Lambda (кількість батьків);

– GA_ProbC (імовірність схрещування) та GA_ProbM (імовірність мутації);

– GA_N (кількість поколінь) та GA_NG (кількість запусків ЕА);

– GA_TournSize (розмір турніру у відборі);

– GA_Gamma (поріг відносної достовірності) та GA_Kappa (поріг носія);

– сутність AGRM відповідає допоміжному ЦПМ, містить первинний ключ AGRM_Date (дата створення) і обов'язкові атрибути значення допоміжного AGR_Data та чіткого допоміжного ЦПМ AGR_CrispData у вигляді текстового рядку з розділеними комами значеннями;

– сутність MMTT_AGRM відповідає викидам допоміжного сигналу: необов'язковий атрибут AGR_Outliers містить індекси значень допоміжного ЦПМ, які відповідають викидам, виявленим з використанням ММТТ у форматі, аналогічному формату зберігання значень самого допоміжного ЦПМ;

– сутність Aux_Visual_Outlier відповідає викидам допоміжного ЦПМ, відібраним статистиком для маскування, містить первинний ключ AVO_Date (дата створення) і обов'язковий атрибут AVO_Outlier з індексами значень допоміжного ЦПМ, які відповідають цим викидам;

– сутність MicrofileProblem потрібна для організації зв'язку «багато-до-багатьох» між мікрофайлами та ЗЗГА.

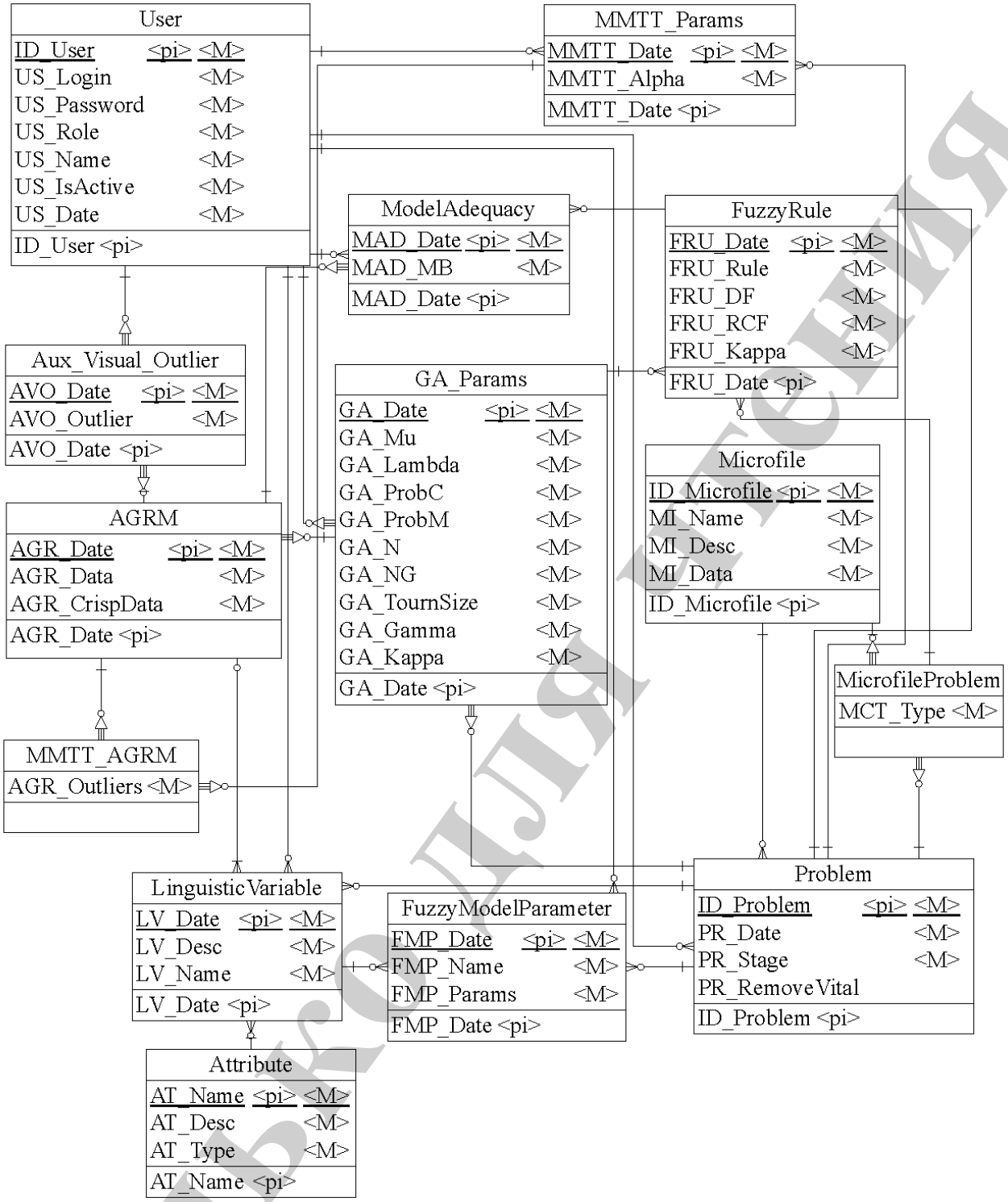

Рис. 5. Фрагмент концептуальної моделі даних, що відповідає нечітким моделям груп у мікрофайлі

На рис. 6 наведено фрагмент концептуальної моделі даних, який відповідає розв’язанню ЗЗГА.

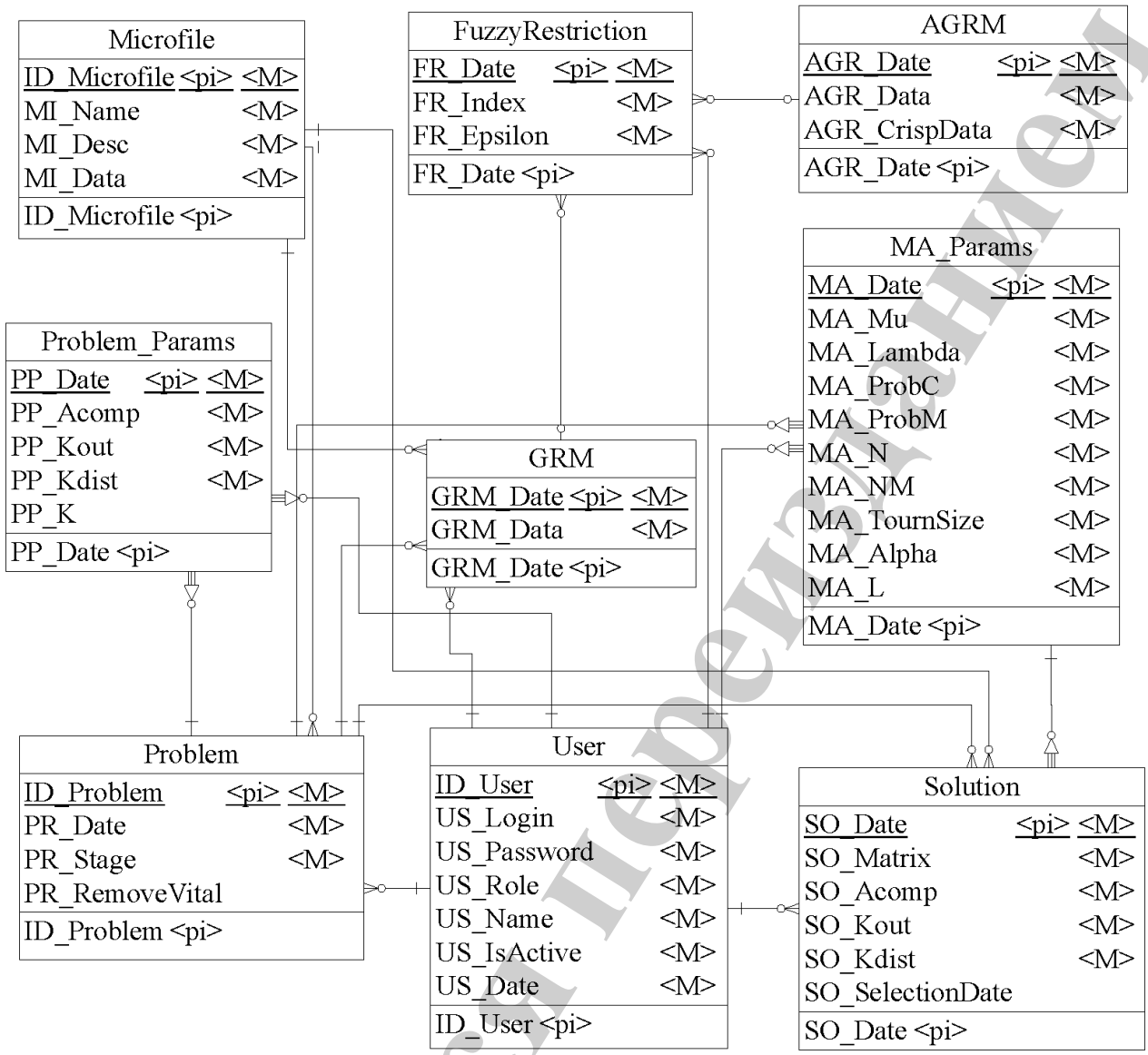

Рис. 6. Фрагмент концептуальної моделі даних, що відповідає розв'язанню ЗЗГА

Сутності, винесені в цей фрагмент, відповідають таким поняттям, пов'язаним із забезпеченням групової анонімності:

– сутність MA_Params відповідає параметрам МА для розв'язання ЗЗГА та містить, крім первинного ключа MA_Date (дата створення параметрів), такі обов'язкові атрибути:

– MA_Mu (розмір популяції) та MA_Lambda (кількість батьків);

– MA_ProbC (імовірність схрещування) та MA_ProbM (імовірність мутації);

– MA_N (кількість поколінь) та MA_NG (кількість запусків МА);

– MA_TournSize (розмір турніру у відборі);

– MA_Alpha (ступінь значущості для ММТТ) та MA_L (максимальна кількість рядків в особині з популяції);

– сутність FuzzyRestriction відповідає нечітким обмеженням на значення ЦПМ та містить, крім первинного ключа FR_Date (дата створення параметрів),

обов’язкові атрибути FR_Index (індекс значення ЦПМ, на яке накладається обмеження) та FR_Epsilon (порогове значення);

– сутність Problem_Params відповідає параметрам ЗЗГА та містить, крім первинного ключа PP_Date (дата створення параметрів), такі обов’язкові атрибути:

– PP_Acomp (поріг сумісності розв’язку ЗЗГА з нечіткими обмеженнями);

– PP_Kout (поріг чутливості розв’язку ЗЗГА);

– PP_Kdist (поріг спотворень розв’язку ЗЗГА) та необов’язковий атрибут PP_K (значення K для формування нечітких обмежень);

– сутність Solution відповідає розв’язкам ЗЗГА та містить, крім первинного ключа SO_Date (дата створення розв’язку), такі обов’язкові атрибути:

– SO_Matrix (розв’язок як особина в МА у вигляді рядку з розділеними комами значеннями);

– SO_Acomp (сумісність розв’язку ЗЗГА з нечіткими обмеженнями);

– SO_Kout (чутливість розв’язку ЗЗГА);

– SO_Kdist (спотворення, внесені розв’язком ЗЗГА) та необов’язковий атрибут SO_SelectionDate (дата відбору розв’язку ЗЗГА).

4. 3. Реалізація інформаційної технології забезпечення групової анонімності

Для реалізації ІТ було вибрано інструментальні засоби, які задовольняють раніше висунуті вимоги:

– як сервер застосунків було вибрано сервер Oracle GlassFish Server, оскільки він забезпечує взаємодію з клієнтами з використанням малої кількості програмних засобів, дає змогу ефективно організовувати роботу з БД та поширюється у вільному доступі. При цьому взаємодія клієнтів із сервером організована за допомогою інтерфейсу Java Message Service, реалізованого за допомогою Apache ActiveMQ, оскільки він дає змогу забезпечити асинхронний обмін повідомленням між сервером та клієнтами, що підвищує гнучкість ІТ;

– як сервер БД було вибрано сервер MySQL, оскільки він простий у застосуванні та поширюється у вільному доступі. При цьому взаємодія клієнтів із БД організована за допомогою інтерфейсу Java Database Connectivity;

– для реалізації застосунків в ІТ було використано дві різні системи:

– платформу Java Platform, Enterprise Edition 8, яка є відносно простою у використанні, портативною, стабільною та поширюється у вільному доступі;

– систему інженерних розрахунків Scilab, яка підтримує матричні обчислення, потрібні для ефективної реалізації методів забезпечення групової анонімності, та на відміну від своїх аналогів поширюється у вільному доступі;

– для реалізації АРМ клієнтів було використано платформу Java Platform, Standard Edition 8 та бібліотеку Swing.

Окремі застосунки ІТ виконують такі функції:

– застосунок створення ЦПМ запускається молодшим аналітиком зі свого АРМ. Застосунок викликає функцію Scilab buildGRM, на вхід якої подаються зчитані з БД дані мікрофайлу та індекси сутнісних і параметризуючих атрибутів із їхніми значеннями. Функція повертає одновимірний масив, кожний елемент якого відповідає кількості респондентів, що належать групі та мають відповідне параметризуюче значення. Застосунок записує відповідний масив у БД та надсилає його в АРМ молодшого аналітика;

– застосунок гармонізації мікрофайлів запускається датологом. Застосунок викликає функцію Scilab harmonize, на вхід якої подаються зчитані з БД дані основного та допоміжного мікрофайлів, а також параметри гармонізації, задані датологом у своєму АРМ. Параметри гармонізації представляються у вигляді об'єкта класу HarmonizationParams. Функція harmonize повертає гармонізовані дані мікрофайлів. Застосунок записує відповідні дані до БД та надсилає в АРМ статистика і датолога гармонізовані метадані обох мікрофайлів;

– застосунок виявлення викидів запускається молодшим аналітиком. Застосунок викликає функцію Scilab detectOutliers, на вхід якої подаються зчитані з БД параметр α ММТТ та ЦПМ. Функція detectOutliers повертає масив обчислених методом ММТТ індексів значень ЦПМ, які є викидами. Застосунок записує відповідний масив до БД та надсилає його в АРМ молодшого аналітика;

– застосунок побудови нечітких правил запускається молодшим аналітиком. Застосунок викликає функцію Scilab ga, на вхід якої подаються зчитані з БД дані допоміжного мікрофайлу, параметри ГА та значення лінгвістичних змінних. Функція виконує ГА для побудови нечіткої моделі та повертає матрицю, рядки якої є нечіткими правилами, а також характеристики цих правил. Застосунок записує знайдені правила до БД та надсилає їх в АРМ молодшого та старшого аналітиків;

– застосунок верифікації адекватності моделі запускає молодший аналітик. Застосунок викликає функцію Scilab getModelAdequacy, на вхід якої подаються зчитані з БД дані про викиди ЦПМ та допоміжного ЦПМ. Функція розраховує значення метрики MB (11). Застосунок записує обчислене значення до БД та надсилає його в АРМ молодшого аналітика;

– застосунок розв'язання ЗЗГА запускається молодшим аналітиком. Застосунок запускає функцію Scilab ma, на вхід якої подаються зчитані з БД дані мікрофайлу, дані ЦПМ, параметри ЗЗГА та ЕА, а також нечіткі обмеження на значення ЦПМ. Функція виконує МА розв'язання ЗЗГА та повертає отримані розв'язки. Застосунок записує розв'язки та їхні характеристики до БД та надсилає в АРМ молодшого і старшого аналітиків. В АРМ відповідних аналітиків розв'язки застосовуються до ЦПМ та на екрані відображаються модифіковані ЦПМ.

4. 4. Опис експериментального дослідження інформаційної технології забезпечення групової анонімності

Для ілюстрації роботи інформаційної технології забезпечення групової анонімності даних розглянемо задачу маскування територіального розподілу військових США у мікрофайлі **M** одновідсоткової вибірки Спостереження за американ-

ським суспільством (2013 р.) [35]. Розв'язання задачі здійснювалося колективом із п'яти фахівців у ролях статистика, датолога, молодшого та старшого аналітиків, а також адміністратора БД.

Мікрофайл **M** містить 1 380 924 записи та, серед інших, такі атрибути:

– «Професія» («Occupation, SOC classification»), значення якого 551010, 552010, 553010 та 559830 (коди професій за Стандартною класифікацією професій США) відповідають військовим різного рангу. Розглядатимемо атрибут «Професія» як сутнісний для даної ЗЗГА;

– «Місце роботи: штат» («Place of work: state, 1980 onward») і «Місце роботи: область мікроданих вільного користування» («Place of work: PUMA, 2000 onward»), які в поєднанні визначають унікальний код територіальної одиниці, у якій працює той чи інший респондент. Розглядатимемо поєднання цих двох атрибутів як параметризуючий атрибут для даної ЗЗГА.

Нехай користувач – старший аналітик вирішив вилучити атрибут «Професія» з мікрофайлу. Основна задача застосування розробленої ІТ полягає в перевірці, чи достатньо такого вилучення для забезпечення анонімності групи військових, і якщо недостатньо, то застосувати відповідний МА.

Як допоміжний мікрофайл $\mathbf{M}^{aux}$ користувач-статистик вибрав мікрофайл п'ятивідсоткової вибірки перепису населення США (2000 р.) [35], який містить 6 309 848 записів. Цей мікрофайл є близький за структурою до основного мікрофайлу **M**, зокрема, обидва мікрофайли було гармонізовано користувачем-датологом у такий спосіб:

– в обох мікрофайлах залишено тільки параметризуючий атрибут «Місце роботи», сутнісний атрибут «Професія» та 13 базових атрибутів, наведених у табл. 2. Кожний базовий атрибут для цілей підрахунку метрики (1) вважається категорійним із вагою 1: метрика (1) показує кількість значень атрибутів, які спотворюються одним обміном записами;

– значення атрибута «Професія» в обох мікрофайлах було трансформовано таким чином: записи, які мали значення 551010, 552010, 553010 та 559830, набули значення «1», решта записів набули значень «0».

Таблиця 2

Базові гармонізовані атрибути для ЗЗГА

№ з/п	Назва українською	Назва англійською	Значення
1	Вік	Age	000 – менше 1 року, від 1 до 130 – вік від 1 до 130 років, 135 – 135 років
2	Стать	Sex	1 – чоловік, 2 – жінка
3	Рівень освіти	Educational attainment [general version]	00 – не застосовується або без освіти, 01 – початкова школа до 4 класу, 02 – 5–8-ий класи, 03 – 9-ий клас, 04 – 10-ий клас, 05 – 11-ий клас, 06 – 12-ий клас, 07 – 1-ий курс ЗВО, 08 – 2-ий курс ЗВО, 09 – 3-ій курс ЗВО, 10 – 4-ий курс ЗВО, 11 – 5-ий курс ЗВО і вище

4	Сімейний стан	Marital status	1 – одружений, проживає з подружжям, 2 – одружений, подружжя проживає окремо, 3 – розлучений, 4 – розірваний шлюб, 5 – удівець, 6 – ніколи не був одружений
5	Сукупний дохід	Total personal income	Дохід респондента за попередній рік у доларах США
6	Кількість робочих годин на тиждень	Usual hours worked per week	00 – не застосовується, від 01 до 98 – від 1 до 98 годин на тиждень, 99 – 99 годин і більше
7	Кількість робочих тижнів на рік (інтервал)	Weeks worked last year, intervalled	0 – не застосовується, 1 – 1–13 тижнів, 2 – 14–26 тижнів, 3 – 27–39 тижнів, 4 – 40–47 тижнів, 5 – 48–49 тижнів, 6 – 50–52 тижнів
8	Раса	Race [general version]	1 – європеоїд, 2 – негроїд, 3 – індіанець, 4 – китаєць, 5 – японець, 6 – інший монголоїд, 7 – інша раса, 8 – дві основні раси, 9 – три і більше основних рас
9	Латиноамериканське походження	Hispanic origin [general version]	0 – не латиноамериканського походження, 1 – мексиканець, 2 – пуерторіканець, 3 – кубинець, 4 – інше, 9 – не вказано
10	Спосіб добирання на роботу	Means of transportation to work	00 – не застосовується, 10 – автомобільний транспорт, 11 – автомобіль, 12 – водій, 13 – пасажир, 14 – грузовик, 15 – фургон, 20 – мотоцикл, 30 – громадський транспорт, 31 – автобус чи тролейбус, 32 – трамвай, 33 – метро, 34 – залізниця, 35 – таксі, 36 – пором, 40 – велосипед, 50 – пішки, 60 – інше, 70 – працює вдома
11	Час виходу на роботу	Time of departure for work	0000 – не застосовується, інші значення відповідають часу виходу на роботу минулого тижня (значення від 0001 до 2359 відповідають моментам часу від 00:01 до 23:59, відповідно)
12	Час на дорогу на роботу	Travel time to work	000 – не застосовується, інші значення відповідають тривалості часу, у хв., який займає дорога на роботу
13	Володіння англійською	Speaks English	0 – не застосовується, 1 – не володіє, 2 – володіє, 3 – володіє тільки англійською, 4 – володіє дуже добре, 5 – володіє добре, 6 – володіє не дуже добре, 7 – невідомо, 8 – неможливо визначити

Лінгвістичні змінні для нечіткої моделі групи L_j, j=1,…, 13, які відповідають базовим гармонізованим атрибутам із таблиці 1, мають параметри, наведені в таблиці 3. У таблиці використовуються такі типи функцій належності:

$$PIMF\left(x;a;b;c;d\right)=$$

$$=\begin{cases}0, & x\le a,\\ 2\left(\dfrac{x-a}{b-a}\right)^2, & a\le x\le \dfrac{a+b}{2},\\ 1-2\left(\dfrac{x-b}{b-a}\right)^2, & \dfrac{a+b}{2}\le x\le b,\\ 1, & b\le x\le c,\\ 1-2\left(\dfrac{x-c}{d-c}\right)^2, & c\le x\le \dfrac{c+d}{2},\\ 2\left(\dfrac{x-d}{d-c}\right)^2, & \dfrac{c+d}{2}\le x\le d,\\ 0, & x\ge d,\end{cases}$$

$$TRAPMF\left(x;a;b;c;d\right)=$$

$$=\begin{cases}0, & x\le a,\\ \dfrac{x-a}{b-a}, & a\le x\le b,\\ 1, & b\le x\le c,\\ \dfrac{d-x}{d-c}, & c\le x\le d,\\ 0, & x\ge d,\end{cases}$$

$$GAUSSMF\left(x;a;b\right)=e^{-\frac{(x-b)^2}{2a^2}},$$

$$SIGMF\left(x;a;b\right)=\frac{1}{1+e^{-a(x-b)}}.$$

Таблиця 3

Параметри лінгвістичних змінних для нечіткої моделі групи для ЗЗГА

№ з/п	Носій базової змінної	Значення
1	[18, 45]	«Дуже молодий»: $\mu_{1,1}(x)$=*PIMF*(*x*; 7,05; 15,40; 22,50; 27,20) «Молодий»: $\mu_{1,2}(x)$=*GAUSSMF*(*x*; 2,0; 27,5) «Середнього віку»: $\mu_{1,3}(x)$=*GAUSSMF*(*x*; 2,0; 32,5) «Не дуже старий»: $\mu_{1,4}(x)$=*GAUSSMF*(*x*; 2,0; 37,5) «Старий»: $\mu_{1,5}(x)$=*PIMF*(*x*; 37,85; 42,50; 47,50; 54,85)
2	[1, 2]	«Чоловік»: $\mu_{2,1}(x)$=*TRAPMF*(*x*;1;1;1;1) «Жінка»: $\mu_{2,2}(x)$=*TRAPMF*(*x*;2;2;2;2)
3	[1, 11]	«Низький»: $\mu_{3,1}(x)$=*TRAPMF*(*x*; 1; 1; 8; 10) «Високий»: $\mu_{3,2}(x)$=*TRAPMF*(*x*; 8; 10; 11;11)
4	[1, 6]	«У шлюбі»: $\mu_{4,1}(x)$=*TRAPMF*(*x*;1; 1; 2; 2) «Поза шлюбом»: $\mu_{4,2}(x)$=*TRAPMF*(*x*; 3; 3; 6; 6)
5	[0, 200000]	«Низький»: $\mu_{5,1}(x)$=*PIMF*(*x*; 0; 0; 9000; 12000) «Середній»: $\mu_{5,2}(x)$=*PIMF*(*x*; 9000; 12000; 70000; 90000) «Високий»: $\mu_{5,3}(x)$=*PIMF*(*x*; 70000; 90000; 200000; 200000)
6	[0, 100]	«Низька»: $\mu_{6,1}(x)$=*PIMF*(*x*;0,0;0,0;29,9;40,3) «Середня»: $\mu_{6,2}(x)$=*GAUSSMF*(*x*;2,5;40,0) «Висока»: $\mu_{6,3}(x)$=*PIMF*(*x*;40,2;50,1;100,0;100,0)
7	[1, 6]	«Нестандартна»: $\mu_{7,1}(x)$=*TRAPMF*(*x*; 1; 1; 5; 6) «Стандартна»: $\mu_{7,2}(x)$=*TRAPMF*(*x*; 5; 6; 6; 6)
8	[1, 2]	«Європеоїд»: $\mu_{8,1}(x)$=*TRAPMF*(*x*; 1; 1; 1; 1) «Негроїд»: $\mu_{8,2}(x)$=*TRAPMF*(*x*; 2; 2; 2; 2)
9	[0, 9]	«Ні»: $\mu_{9,1}(x)$=*TRAPMF*(*x*; 0; 0; 0; 0) «Так»: $\mu_{9,2}(x)$=*TRAPMF*(*x*; 1; 1; 9; 9)
10	[0, 70]	«Власний транспорт»: $\mu_{10,1}(x)$=*TRAPMF*(*x*; 0; 0; 20; 20) «Публічний транспорт»: $\mu_{10,2}(x)$=*TRAPMF*(*x*; 30; 30; 36; 36) «Пішки»: $\mu_{10,3}(x)$=*TRAPMF*(*x*; 40; 40; 50; 50)
11	[1, 2359]	«Ніч»: $\mu_{11,1}(x)$=*PIMF*(*x*; 1; 1; 530; 630) «Ранок»: $\mu_{11,2}(x)$=*PIMF*(*x*; 530; 630; 800; 900) «День»: $\mu_{11,3}(x)$=*PIMF*(*x*; 800; 900; 2359; 2359)
12	[1, 119]	«Нетривалий»: $\mu_{12,1}(x)$=*PIMF*(*x*; 1; 1; 10; 15) «Середньої тривалості»: $\mu_{12,2}(x)$=*PIMF*(*x*; 10; 15; 35; 45) «Тривалий»: $\mu_{12,3}(x)$=*PIMF*(*x*; 35; 45; 120; 120)
13	[2, 5]	Немає (змінна використовується тільки для вилучення недопустимих записів із мікрофайлів)

Нечітка модель групи військових будувалася за допомогою ГА з параметрами, наведеними в табл. 4.

Таблиця 4

Параметри генетичного алгоритму для побудови нечіткої моделі групи військових

Параметр	Значення
Розмір популяції μ	100
Кількість пар батьківських особин λ	20
Імовірність схрещування p_c	1,000
Імовірність мутації p_m	0,050
Розмір турніру у відборі	10
Кількість запусків алгоритму	10
Кількість поколінь у кожному запуску	100
Поріг відносної достовірності γ	0,750
Поріг носія κ	0,001

Таким чином, після вилучення з мікрофайлів **M** та $\mathbf{M}^{aux}$ записів, значення яких не належать носіям базових змінних лінгвістичних змінних L_j, j=1,…, 13, основний мікрофайл **M** став містити 565 243 записи, із яких 3 992 – сутнісні, допоміжний мікрофайл $\mathbf{M}^{aux}$ став містити 3 205 478 записів, із яких 14 263 – сутнісні.

5. Результати експерименту з випробування інформаційної технології забезпечення групової анонімності

У результаті застосування ГА з параметрами, наведеними в таблиці 4, було побудовано нечітку модель групи військових, яка складається з правил, наведених разом зі своїми характеристиками в табл. 5. Варто звернути увагу, що в усіх правилах лінгвістична змінна «Спосіб добирання на роботу» представлена значенням «Пішки», що, вочевидь, є характеристичною особливістю військовослужбовців, які проживають у казармах.

Таблиця 5

Правила нечіткої моделі групи військових

Правило	*DF*	*RCF*	κ
(1, 0, 0, 2, 2, 0, 0, 0, 0, 3, 1, 1)	0,032	0,755	0,032
(1, 0, 0, 0, 0, 3, 0, 0, 0, 3, 1, 0)	0,031	0,787	0,031
(1, 0, 0, 0, 1, 3, 2, 0, 1, 3, 0, 0)	0,012	0,801	0,012
(1, 0, 0, 0, 2, 0, 1, 1, 1, 3, 1, 1)	0,010	0,781	0,010
(1, 0, 0, 0, 1, 3, 2, 1, 0, 3, 0, 0)	0,012	0,851	0,012
(1, 1, 0, 0, 2, 0, 0, 0, 0, 3, 1, 1)	0,034	0,840	0,034
(1, 1, 0, 2, 0, 0, 2, 0, 0, 3, 1, 1)	0,025	0,765	0,025
(1, 1, 0, 2, 1, 3, 0, 0, 0, 3, 2, 0)	0,018	0,931	0,018
(1, 1, 0, 0, 1, 3, 0, 0, 1, 3, 2, 0)	0,017	0,915	0,018
(1, 1, 0, 0, 0, 0, 2, 1, 0, 3, 1, 1)	0,025	0,754	0,026
(1, 1, 0, 2, 2, 0, 0, 1, 0, 3, 1, 0)	0,032	0,751	0,032
(1, 0, 1, 2, 1, 3, 0, 0, 0, 3, 2, 1)	0,018	0,951	0,018
(1, 0, 1, 0, 1, 3, 0, 0, 1, 3, 2, 0)	0,019	0,767	0,019

(1, 1, 1, 0, 1, 3, 2, 0, 0, 3, 2, 0)	0,009	1,876	0,009
(1, 1, 1, 0, 1, 3, 1, 0, 0, 3, 2, 1)	0,008	0,761	0,009
(1, 1, 1, 2, 1, 3, 2, 0, 0, 3, 0, 1)	0,010	1,325	0,010
(1, 1, 1, 2, 0, 3, 2, 0, 0, 3, 2, 1)	0,026	0,767	0,026
(1, 1, 2, 0, 2, 0, 0, 0, 0, 3, 1, 0)	0,002	0,914	0,002

Аналізувати застосування побудованої нечіткої моделі до виявлення викидів ЦПМ основного мікрофайлу **M** доцільно робити окремо для кожного штату. Так, зокрема, для штату Нью-Йорк цільове подання мікрофайлу **M** відносно групи військових $\mathbf{q}_{NY}$ та допоміжне цільове подання мікрофайлу $\mathbf{M}^{aux}$ відносно групи військових $\mathbf{q}^{aux}_{NY}$ наведено на рис. 7. Вісь абсцис на рисунку відповідає територіальним одиницям, у яких служать військові (значення 1–33 відповідають значенням параметризуючого атрибута 3600100–3603300, відповідно, значення 34–38 відповідають значенням параметризуючого атрибута 3603700–3604100, відповідно), а вісь ординат – кількостям військових, що там служать.

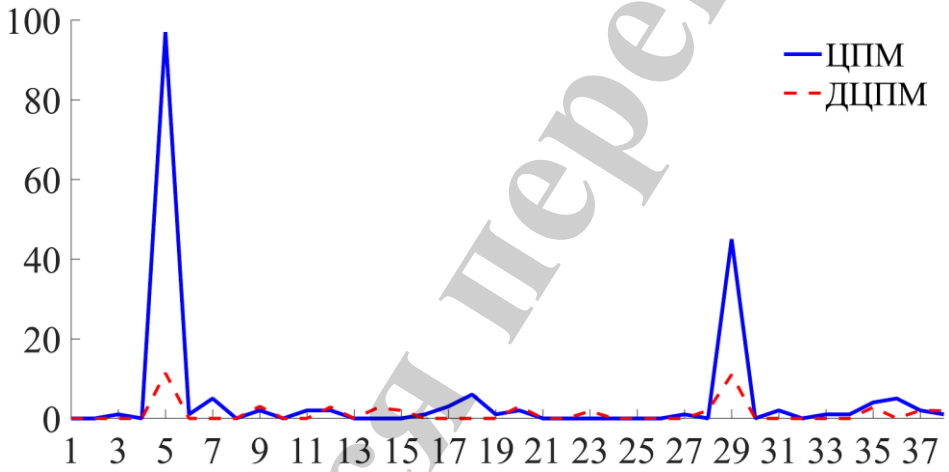

Рис. 7. ЦПМ та допоміжне ЦПМ відносно групи військових

Застосування ММТТ з параметром α=0,01 до наведеного ЦПМ дало змогу користувачу – молодшому аналітику отримати множину індексів $OUT(\mathbf{q}_{NY})$={5, 7, 9, 11, 12, 17, 18, 20, 29, 31, 35, 36, 37}. У цій множині індексів тільки два індекси відповідають справжнім військовим базам [36], тому користувач – молодший аналітик залишив тільки їх для подальшого аналізу: $OUT(\mathbf{q}_{NY})$={5, 29}. Індекс 5 відповідає Форту Драм, а індекс 29 – Військовій академії Вест-Поїнт.

Аналогічні міркування після застосування ММТТ з параметром α=0,01 до допоміжного ЦПМ дали змогу користувачу – молодшому аналітику отримати множину $OUT(\mathbf{q}^{aux}_{NY})$={5, 29}. Рівність двох множин безпосередньо свідчить про можливість порушення групової анонімності групи військових, навіть якщо атрибут «Професія» буде вилучено з мікрофайлу. Аналогічний аналіз для решти штатів США, у яких кількість військових перевищує 0,5 % від загального числа військових у **M**, наведено в табл. 6.

Таблиця 6

Результати застосування нечіткої моделі групи військових до мікрофайлу **M**

Штат	Кількість викидів у ЦПМ	Кількість викидів у ЦПМ, відсутніх у ДЦПМ	Кількість викидів у ДЦПМ	Кількість викидів у ДЦПМ, відсутніх у ЦПМ
Алабама	2	2	1	1
Аляска	2	0	2	0
Аризона	4	1	4	1
Вашингтон	4	1	3	0
Вірджинія	7	4	4	1
Гаваї	1	0	1	0
Джорджія	7	3	4	0
Іллінойс	2	1	2	1
Каліфорнія	3	1	2	0
Канзас	2	2	0	0
Кентуккі	2	1	1	0
Колорадо	2	0	2	0
Коннектикут	1	0	2	1
Луїзіана	4	4	0	0
Мериленд	3	2	1	0
Міссісіпі	1	0	1	0
Міссурі	2	2	0	0
Невада	1	0	1	0
Нью-Джерсі	2	2	0	0
Нью-Мехіко	2	2	0	0
Нью-Йорк	2	0	2	0
Огайо	2	1	3	2
Оклагома	3	2	1	0
Південна Кароліна	4	1	3	0
Північна Кароліна	3	1	2	0
Техас	6	1	5	0
Флорида	7	5	3	1
Загалом	81	39	50	8

Матриця невідповідностей для прикладу дорівнює

$$\mathbf{Z} = \begin{pmatrix} 48 & 39 \\ 8 & 564 \end{pmatrix}.$$

На основі цієї матриці користувач – молодший аналітик підрахував метрику адекватності нечіткої моделі (11), яка дорівнює *MB*=55,067, що свідчить про високу адекватність моделі.

Для надійного забезпечення групової анонімності, таким чином, недостатньо просто вилучити з мікрофайлу **M** сутнісний атрибут «Професія», а потрібно додатково застосовувати МА для маскування викидів допоміжного ЦПМ $\mathbf{q}^{aux}$. Розглянемо приклад застосування ІТ для штату Нью-Йорк.

Користувач – молодший аналітик наклав на 5 та 29 відлік ЦПМ нечіткі обмеження з функцією належності

$$Z_{MF}\left(x,a,b\right)=\begin{cases}1, & x\le a,\\ 1-2\left(\dfrac{x-a}{b-a}\right)^2, & a\le x\le \dfrac{a+b}{2},\\ 2\left(\dfrac{x-b}{b-a}\right)^2, & \dfrac{a+b}{2}\le x\le b,\\ 0, & x\ge b\end{cases}$$

і функція пристосованості (5) для прикладу набула вигляду:

$$f\left(U\right)=\frac{299-\sum\limits_{i=1}^{Q}\sum\limits_{k=1}^{13}\operatorname{sign}\left|\mathbf{M}_{u_{i1}}\left(u_{i2},w_{b_k}\right)-\mathbf{M}_{u_{i3}}\left(u_{i4},w_{b_k}\right)\right|}{299}\times$$

$$\times ZMF\left(q_{NY_5}^{aux*}\left(U\right),2,12\right)\cdot ZMF\left(q_{NY_{29}}^{aux*}\left(U\right),2,11\right)\times$$

$$\times\frac{1}{1+e^{\frac{1}{2}(Q-25)}},$$

де $C_{\max}$=299; w_{bk} – *k*-ий базовий атрибут, *k*=1,…, 13.

Користувач – молодший аналітик підібрав параметри МА для маскування викидів у ДЦПМ, наведені в табл. 7. При цьому ймовірність мутації збільшувалася в 10 разів, коли середньоквадратичне відхилення значень функції пристосованості особин у деякій популяції ставало меншим за 0,03.

Таблиця 7

Параметри міметичного алгоритму розв’язання ЗЗГА

Параметр	Значення
Розмір популяції μ	100
Кількість пар батьківських особин λ	20
Імовірність схрещування p_c	1,000
Імовірність мутації p_m	0,001
Параметр локального пошуку p_{mem}	0,750

Розмір турніру у відборі	5
Поріг сумісності α_{comp}	0,500
Поріг чутливості K_{out}	0,000
Поріг спотворень K_{dist}	0,250
Кількість запусків алгоритму	10
Кількість поколінь у кожному запуску	1000

Результатами роботи МА із зазначеними параметрами стали 1000 особин останніх поколінь кожного запуску, із яких 983 відповідають вимоги щодо порогів сумісності, чутливості та спотворень. Середнє значення метрики (1) по всіх 983 розв'язках дорівнює 62,518, тобто забезпечення анонімності досягається шляхом спотворення 62,518/(13·1380924)≈0,0003 % значень атрибутів мікрофайлу.

Розв'язок $\mathbf{q}^{aux}*$ із найменшим значенням, що дорівнює 53, метрики (1), подано на рис. 8.

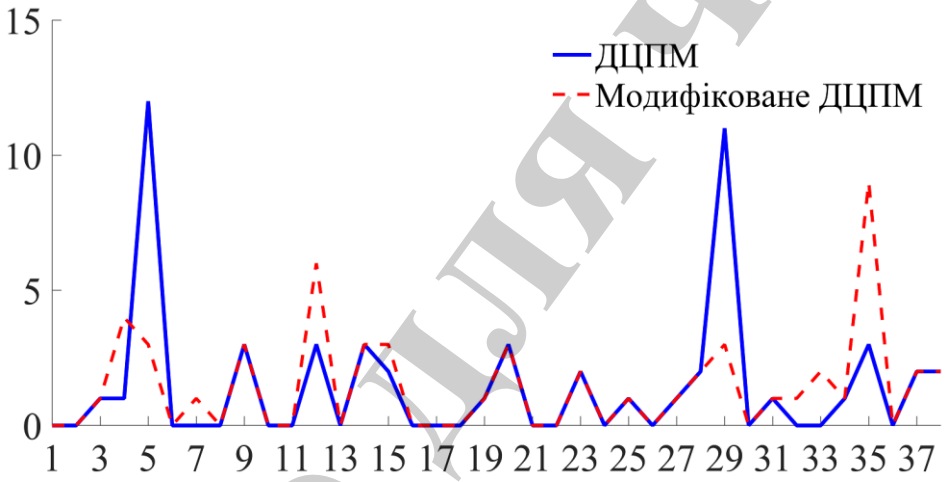

Рис. 8. Допоміжне ЦПМ $\mathbf{q}^{aux}$ та модифіковане допоміжне ЦПМ $\mathbf{q}^{aux}*$

Викиди в $\mathbf{q}^{aux}*$, одержані за допомогою ММТТ, відповідають елементам з індексами 12 та 35, тобто $OUT(\mathbf{q}^{aux})\cap OUT(\mathbf{q}^{aux}*)=\emptyset$.

6. Обговорення результатів розв'язання ЗЗГА за допомогою інформаційної технології

У наведеному експериментальному дослідженні продемонстровано як можна застосувати запропоновану інформаційну технологію забезпечення групової анонімності даних в автоматизованому режимі. При цьому показано, що розроблена ІТ задовольняє висунуті вище вимоги, оскільки вона:

– дає змогу будувати моделі груп респондентів у мікрофайлі за рахунок задання параметризуючих та сутнісних атрибутів та їхніх значень;

– дає змогу будувати нечіткі моделі груп респондентів у мікрофайлі на основі генетичних алгоритмів, параметри яких гнучко задаються;

– дає змогу забезпечити групову анонімність даних за допомогою міметичного алгоритму, який вносить у дані спотворення малого обсягу.

При цьому різні операції та дії в рамках процесу забезпечення групової анонімності даних виконуються користувачами з різними ролями, що дає змогу підвищити ефективність підготовки мікрофайлів до опублікування за рахунок розподілу праці та спеціалізації окремих фахівців. До таких операцій належать гармонізація мікрофайлів, підбір допоміжного мікрофайлу, параметризація алгоритмів та методів забезпечення групової анонімності, ухвалення рішення про вилучення сутнісних атрибутів та остаточне завершення процесу анонімізації даних.

Високий рівень надійності та безпеки первинних даних забезпечується об'єднанням усіх складових ІТ в локальну мережу з обмеженим доступом.

Розв'язання задачі з експерименту колективом із п'яти фахівців зайняло 7 год 50 хв., у той час, як розв'язання аналогічної задачі за допомогою ІТ, описаної в [20] – 19 год 45 хв, тобто в 2,5 рази довше. Таке підвищення швидкості підготовки мікрофайлів до оприлюднення пояснюється організацією ефективної взаємодії користувачів різного фаху та інтеграції в ІТ методів, які забезпечують більш ефективне розв'язання ЗЗГА порівняно з описаними в [20].

Додаткових досліджень потребує розроблення інструкції щодо найбільш ефективного використання ІТ, зокрема, рекомендацій користувачу – молодшому аналітику щодо підбору параметрів ГА та МА та критеріїв завершення процесу анонімізації даних для користувача – старшого аналітика. Наявність такої інструкції суттєво розширить коло користувачів ІТ та уможливить підготовку даних до оприлюднення організаціями, які не спеціалізуються на їх статистичній обробці.

7. Висновки

1. Запропоновано трирівневу клієнт-серверну архітектуру інформаційної технології забезпечення групової анонімності даних, у якій виділено клієнтів, сервери застосунків та бази даних, об'єднані в локальну мережу. Технологія враховує можливість порушення групової анонімності в умовах доступу до допоміжного мікрофайлу, що дає змогу підвищити рівень захищеності даних.

2. Розроблено концептуальну модель реляційної БД для запропонованої ІТ, яка містить усі сутності процесу забезпечення групової анонімності даних та відображає зв'язки між ними. Наведені ключові фрагменти побудованої моделі даних, які відповідають сутностям мікрофайлу та його цільового подання, нечітким моделям груп у мікрофайлі та розв'язанню задачі забезпечення групової анонімності.

3. Розглянуто реалізацію технології на основі платформи Java Enterprise Edition 8, сервера застосунків Oracle GlassFish, сервера БД MySQL та системи інженерних розрахунків SciLab. Описано взаємодію застосунків у технології з АРМ клієнтів та функціями, написаними в системі SciLab. Запропонована реалізація задовольняє вимоги, висунуті до інформаційної технології.

4. Практичне застосування інформаційної технології проілюстровано розв'язанням задачі анонімізації групи військових на основі реальних даних

Спостереження за американським суспільством 2013 р. Встановлено, що застосування технології дає змогу в 2,5 рази пришвидшити процес підготовки мікрофайлу до опублікування силами колективу з п’яти фахівців.

Література

1. Duncan G. T., Elliot M., Salazar-González J.-J. Statistical Confidentiality. Principles and Practice. Springer-Verlag, 2011. 212 p. doi: https://doi.org/10.1007/978-1-4419-7802-8
2. A Terminology for Talking About Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management, Version v0.34. URL: http://dud.inf.tu-dresden.de/Anon_Terminology.shtml
3. Чертов, О. Р., Тавров Д. Ю. Забезпечення групової анонімності як складова CSID-процесу обробки даних // Штучний інтелект. 2017. № 3-4. С. 127–138.
4. Chertov O., Tavrov D. Improving efficiency of providing data group anonymity by automating data modification quality evaluation // Eastern-European Journal of Enterprise Technologies. 2017. Vol. 5, Issue 4 (89). P. 31–39. doi: https://doi.org/10.15587/1729-4061.2017.113046
5. Chertov O., Tavrov D. Microfiles as a Potential Source of Confidential Information Leakage // Studies in Computational Intelligence. 2014. P. 87–114. doi: https://doi.org/10.1007/978-3-319-08624-8_4
6. Tavrov D., Chertov O. Evolutionary approach to violating group anonymity using third-party data // SpringerPlus. 2016. Vol. 5, Issue 1. doi: https://doi.org/10.1186/s40064-016-1692-9
7. Butz M. V. Learning Classifier Systems // Springer Handbook of Computational Intelligence. 2015. P. 961–981. doi: https://doi.org/10.1007/978-3-662-43505-2_47
8. Holland J. H. Adaptation in Natural and Artificial Systems. Ann Arbor, MI: University of Michigan Press, 1975. 183 p.
9. Valenzuela-Rendón M. The Fuzzy Classifier System: Motivations and First Results // Proceedings of Parallel Solving from Nature (PPSN II). 1991. P. 330–334.
10. Smith S. F. A Learning System Based on Genetic Adaptive Algorithms. Pittsburgh: University of Pittsburgh, 1980. 214 p.
11. Overview on evolutionary subgroup discovery: analysis of the suitability and potential of the search performed by evolutionary algorithms / Carmona C. J., González P., del Jesus M. J., Herrera F // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2014. Vol. 4, Issue 2. P. 87–103. doi: https://doi.org/10.1002/widm.1118
12. Ishibuchi H., Nakashima T., Murata T. Performance evaluation of fuzzy classifier systems for multidimensional pattern classification problems // IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics). 1999. Vol. 29, Issue 5. P. 601–618. doi: https://doi.org/10.1109/3477.790443

13. μ-ARGUS Version 5.1.3. User's Manual / Hundepool A., de Wolf P.-P., Bakker J., Reedijk A., Franconi L. et. al. 2018. URL: http://neon.vb.cbs.nl/casc/Software/MUmanual5.1.3.pdf

14. Angiuli O., Waldo J. Statistical Tradeoffs Between Generalization and Suppression in the De-Identification of Large-Scale Data Sets // 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC). 2016. doi: https://doi.org/10.1109/compsac.2016.198

15. Sweeney L. K-Anonymity: A Model for Protecting Privacy // International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002. Vol. 10, Issue 05. P. 557–570. doi: https://doi.org/10.1142/s0218488502001648

16. Fienberg S., McIntyre J. Data Swapping: Variations on a Theme by Dalenius and Reiss // Journal of Official Statistics. 2005. Vol. 21, Issue 2. P. 309–323.

17. Evfimievski A. Randomization in privacy preserving data mining // ACM SIGKDD Explorations Newsletter. 2002. Vol. 4, Issue 2. P. 43–48. doi: https://doi.org/10.1145/772862.772869

18. Templ M. Statistical Disclosure Control for Microdata Using the R-package sdcMicro // Transactions on Data Privacy. 2008. Vol. 1, Issue 2. P. 67–85.

19. Domingo-Ferrer J., Mateo-Sanz J. M. Practical data-oriented microaggregation for statistical disclosure control // IEEE Transactions on Knowledge and Data Engineering. 2002. Vol. 14, Issue 1. P. 189–201. doi: https://doi.org/10.1109/69.979982

20. Чертов О. Р. Мінімізація спотворень при формуванні мікрофайлу з замаскованими даними // Вісник Східноукраїнського національного університету ім. В. Даля. 2012. № 8 (179). С. 240–246.

21. Chertov O., Tavrov D. Providing Group Anonymity Using Wavelet Transform // Lecture Notes in Computer Science. 2012. P. 25–36. doi: https://doi.org/10.1007/978-3-642-25704-9_5

22. Chertov O., Tavrov D. Two-Phase Memetic Modifying Transformation for Solving the Task of Providing Group Anonymity // Studies in Fuzziness and Soft Computing. 2016. P. 239–253. doi: https://doi.org/10.1007/978-3-319-32229-2_17

23. Zadeh L. A. Toward a restriction-centered theory of truth and meaning (RCT) // Information Sciences. 2013. Vol. 248. P. 1–14. doi: https://doi.org/10.1016/j.ins.2013.06.003

24. Neri F., Cotta C. A Primer on Memetic Algorithms // Studies in Computational Intelligence. 2012. P. 43–52. doi: https://doi.org/10.1007/978-3-642-23247-3_4

25. Goldberg D. E., Korb B., Deb K. Messy Genetic Algorithms: Motivation, Analysis, and First Results // Complex Systems. 1989. Vol. 3. P. 493–530.

26. Syswerda G. Schedule Optimization Using Genetic Algorithms // Handbook of Genetic Algorithms. New York: Van Nostrand Reinhold, 1991. P. 332–349.

27. Eiben A. E., Smith J. E. Introduction to Evolutionary Computing. Springer-Verlag, 2015. 287 p. doi: https://doi.org/10.1007/978-3-662-44874-8

28. Brindle A. Genetic Algorithms for Function Optimization. Edmonton: University of Alberta, 1981. 193 p.

29. Zadeh L. A. The concept of a linguistic variable and its application to approximate reasoning – I // Information Sciences. 1975. Vol. 8, Issue 3. P. 199–249. doi: https://doi.org/10.1016/0020-0255(75)90036-5

30. Wrobel S. An algorithm for multi-relational discovery of subgroups // Lecture Notes in Computer Science. 1997. P. 78–87. doi: https://doi.org/10.1007/3-540-63223-9_108

31. Lavrač N., Flach P., Zupan B. Rule Evaluation Measures: A Unifying View // Lecture Notes in Computer Science. 1999. P. 174–185. doi: https://doi.org/10.1007/3-540-48751-4_17

32. Selecting fuzzy if-then rules for classification problems using genetic algorithms / Ishibuchi H., Nozaki K., Yamamoto N., Tanaka H. // IEEE Transactions on Fuzzy Systems. 1995. Vol. 3, Issue 3. P. 260–270. doi: https://doi.org/10.1109/91.413232

33. Syswerda G. Uniform Crossover in Genetic Algorithms // Proceedings of the 3rd International Conference on Genetic Algorithms. Morgan Kaufmann Publishers Inc., 1989. P. 2–9.

34. Olivetti E., Greiner S., Avesani P. Statistical independence for the evaluation of classifier-based diagnosis // Brain Informatics. 2014. Vol. 2, Issue 1. P. 13–19. doi: https://doi.org/10.1007/s40708-014-0007-6

35. Integrated Public Use Microdata Series, Version 8.0 [Dataset] / Ruggles S., Flood S., Goeken R., Grover J., Meyer E., Pacas J., Sobek M. Minneapolis: University of Minnesota, 2018. URL: https://usa.ipums.org/usa/

36. 2011 Demographics. Profile of the Military Community. 2012. URL: http://www.militaryonesource.mil/12038/MOS/Reports/2011_Demographics_Report.pdf